University of Zurich

Department of Economics

Working Paper Series

ISSN 1664-7041 (print)
ISSN 1664-705X (online)

Working Paper No. 335

Quadratic Shrinkage for Large Covariance Matrices

Olivier Ledoit and Michael Wolf

Revised version, December 2020

Quadratic Shrinkage for Large Covariance Matrices*

Olivier Ledoit
Department of Economics
University of Zurich
CH-8032 Zurich, Switzerland
olivier.ledoit@econ.uzh.ch

Michael Wolf
Department of Economics
University of Zurich
CH-8032 Zurich, Switzerland
michael.wolf@econ.uzh.ch

December 2020

Abstract

This paper constructs a new estimator for large covariance matrices by drawing a bridge between the classic Stein (1975) estimator in finite samples and recent progress under large-dimensional asymptotics. The estimator keeps the eigenvectors of the sample covariance matrix and applies shrinkage to the *inverse* sample eigenvalues. The corresponding formula is *quadratic*: it has two shrinkage targets weighted by quadratic functions of the concentration (that is, matrix dimension divided by sample size). The first target dominates mid-level concentrations and the second one higher levels. This extra degree of freedom enables us to outperform linear shrinkage when optimal shrinkage is not linear (which is the general case). Both of our targets are based on what we term the "Stein shrinker", a local attraction operator that pulls sample covariance matrix eigenvalues towards their nearest neighbors, but whose force diminishes with distance, like gravitation. We prove that no cubic or higher-order nonlinearities beat quadratic with respect to Frobenius loss under large-dimensional asymptotics. Non-normality and the case where the matrix dimension exceeds the sample size are accommodated. Monte Carlo simulations confirm state-of-the-art performance in terms of accuracy, speed, and scalability.

KEY WORDS: Inverse shrinkage, Hilbert transform, large-dimensional asymptotics, signal amplitude, Stein shrinkage.

JEL CLASSIFICATION NOS: C13.

*We thank an Associate Editor and two anonymous referees who contributed to the enhancement of the paper. We also thank Patrick Ledoit for acting in the capacity of a research assistant by writing the Python version of our Matlab programming code in Appendix A.

1 Introduction

The covariance matrix is, arguably, the second most important object in all of statistics. It has long been known — by theoreticians and practitioners alike — that the sample covariance matrix suffers from the curse of dimensionality. This curse is most obvious when the matrix dimension exceeds the sample size — in which case the sample covariance matrix is singular — but is pervasive unless matrix dimension is *negligible* with respect to sample size.

Efforts to robustify covariance matrix estimation against large dimensions can be broadly divided into two generations. First, the 20th century, characterized by (1.a) finite-sample mathematics, (1.b) the normality assumption, and (1.c) matrix dimensions below the sample size. Second, the 21st century, characterized by (2.a) large-dimensional asymptotics, (2.b) relaxation of the normality assumption, and (2.c) matrix dimensions below or above the sample size. At the risk of over-simplification, the second generation can be summarized as giving first-generation (finite-sample) ideas a major upgrade thanks to the powerful mathematics of large-dimensional asymptotics. For example, Ledoit and Wolf (2004) apply linear shrinkage to the covariance matrix in the spirit of James and Stein (1961). Bodnar et al. (2016) linearly shrink the precision matrix (inverse of the covariance matrix) as Haff (1979) did. The tests for identity and sphericity of the covariance matrix proposed by Ledoit and Wolf (2002) and Chen et al. (2010) are direct heirs of the ones of John (1971) and Nagao (1973).

The crowning achievement of the first generation of research on robustifying covariance matrix estimation against the curse of dimensionality is without doubt the nonlinear shrinkage formula of Stein (1975, 1977, 1986). A string of Monte Carlo simulations starting with Lin and Perlman (1985) have found it remarkably accurate, especially when the cross-sectional distribution of covariance matrix eigenvalues (a.k.a. principal components) is not smooth but clustered, a difficult case to handle. According to Rajaratnam and Vincenzi (2016b), "Stein's covariance estimator is considered a gold standard in the literature". Yet the Stein shrinkage estimator has not been given its rightful 'large-dimensional asymptotic upgrade' yet. This is the objective of the present paper.

We start from a solid foundation in finite samples by reinterpreting Stein's highly nonlinear (and not immediately intuitive) formula as just linear shrinkage in inverse-eigenvalues space. This gives much greater clarity and insight into what is really happening under the hood. The linear shrinkage intensity could not be simpler: it is the "concentration (ratio)", defined as the ratio of matrix dimension to sample size, a standard measure of the severity of the curse of dimensionality. The higher the concentration, the more the eigenvalues need to be shrunk away from the observed ones. This is called "shrinkage" because the cross-sectional dispersion of the eigenvalues goes down, as they are attracted to one another. What is more interesting is that the shrinkage target is not always the same: it varies depending on relative position with respect to surrounding sample eigenvalues. An eigenvalue that lies slightly above (below) a concentrated cluster of the other eigenvalues is attracted downwards (upwards), and the intensity of this attraction vanishes as the distance increases. Stein's results, reinterpreted in this light, provide

a very well-defined targeting function that captures this important phenomenon, and we call this function the "Stein shrinker". It will be the central object throughout the paper.

The only problem with the naïve Stein shrinker is that it explodes when two consecutive eigenvalues get too close to each other. This is where large-dimensional asymptotics comes into play. We show that smoothing the Stein shrinker provides a covariance matrix estimator that is optimal with respect to Stein's loss under large-dimensional asymptotics. The smoothing parameter must vanish asymptotically as the matrix dimension and the sample size go to infinity together, but not too fast. Note that Stein himself, even though he had essentially the same formula (up to smoothing), could not formally prove optimality in finite samples. Note also that Stein (1986) explicitly acknowlegded both the need for injecting some type of smoothing (*ex-post* through his so-called isotonization algorithm), and also the relevance of large-dimensional asymptotics (cf. his Theorem 1). Therefore, the "smoothed Stein shrinker", as we call it, is just a reorganization of the fundamental ingredients that were already embedded in Stein's original work, but could not be brought to full fruition at that time.

These developments naturally open the door to other loss functions. The first obvious candidate is simply the Inverse Stein's loss (Tsukuma, 2005): Stein's loss applied to the precision matrix instead of the covariance matrix. This loss function belongs to a broader class that contains two more loss functions: the Frobenius loss, and the Minimum Variance loss of Engle et al. (2019). The former has proven quite popular in a variety of applied fields ranging from macroeconomics (Korniotis, 2008) to brain-computer interface (Vidaurre et al., 2009) to analytical chemistry (Guo et al., 2012), among many others. The latter is ideal whenever the objective is to optimize the reward-to-risk ratio (finance) or signal-to-noise ratio (electrical engineering). The reason why these three loss functions are grouped together is that they lead to the same optimal nonlinear shrinkage formula, both in finite samples and under large-dimensional asymptotics.

To handle these loss functions, the asymptotically optimal solution is to move from *linear* shrinkage in inverse-eigenvalues space to *quadratic* shrinkage. There are now two shrinkage targets: one driven by the smoothed Stein shrinker as before, and the other by its "squared amplitude". For readers not versed in signal processing, the squared amplitude is basically the square of the Stein shrinker, but with something more added. The extra part is the square of the 'hidden' or imaginary component that is the conjugate of the Stein shrinker. This is a basic concept in signal processing that goes back to Gabor (1946), but we review all the necessary details in the main body of the paper. As for the shrinkage intensities themselves, they are split three-ways between the original inverse eigenvalues, the smoothed Stein shrinker, and its squared amplitude. The first dominates small concentrations, the second dominates for mid-level concentrations around the 0.5 mark, and the third dominates for high concentrations tending to 1. All three shrinkage intensities are quadratic functions of the concentration ratio.

To cap it all off, we venture into the realm where the sample covariance matrix is singular. In this case, the quadratic weighting scheme delineated above degenerates into putting 100% weight on the squared amplitude of the smoothed Stein shrinker. Null sample eigenvalues need to be kicked out of the computation of the shrinker, so instead they get their own shrinkage formula:

a simple function of the concentration ratio and the harmonic mean of non-null eigenvalues.

Rolling out the 'large-dimensional asymptotic upgrade' generates many obvious contributions with respect to Stein (1975, 1986), listed above. With respect to linear shrinkage, the contribution is also clear because going from linear to quadratic is an easily manageable enhancement that guarantees maximum accuracy even when eigenvalues can be dispersed, clustered, or otherwise unruly. With respect to the other nonlinear shrinkage formulas from Ledoit and Wolf (2015) onwards, there are two key advantages. The first is that the formula obtained here comes from classical statistics instead of from random matrix theory (RMT). All existing ones have been based on a fundamental equation from RMT originally due to Marčenko and Pastur (1967), reformulated by Silverstein (1995), and generalized by Ledoit and Péché (2011). By starting from Stein's (1975) first-generation classic instead, we not only reconnect with a rich body of literature in multivariate statistics, but also inject much-needed understandability. Although many results from RMT have been used for quite a while now by some statisticians, the field is, arguably, still not overly familiar to others. By contrast, it is plain to see, just by visually inspecting the Stein shrinker itself, that eigenvalues are attracted to close-by clusters of other eigenvalues, whereas distant clusters have diminishing influence. We hope that this feature makes the resulting estimator more transparent and user-friendly because opacity usually slows down adoption. The second key advantage is that we manage to reduce mathematical complexity from an infinite degree of nonlinearity to just two degrees (quadratic shrinkage). All this is accomplished without sacrificing accuracy, computational speed, or scalability. In particular, concerning accuracy, we attain the same performance in the large-dimensional limit as the best nonlinear shrinkage formulas based on the fundamental equation of RMT, and (almost) the same performance in finite samples as well.

The remainder of the paper is organized as follows. Section 2 reinterprets the classic first-generation Stein (1986) paper as linear shrinkage in inverse-eigenvalues space and introduces what we term the Stein shrinker. Section 3 shows how to smooth out the explosive discontinuity inside the Stein shrinker, and states conditions on the smoothing parameter that guarantee optimality with respect to Stein's loss under large-dimensional asymptotics. Section 4 adapts the formula to the Inverse Stein, Frobenius, and Minimum Variance loss functions by introducing a second shrinkage target based on the squared amplitude of the smoothed Stein shrinker, and by making shrinkage intensities quadratic functions of the concentration ratio. Section 5 shows how to handle the case when the dimension exceeds the sample size. Section 6 conducts an extensive numerical calibration to select a smoothing parameter in the theoretically acceptable range, which results in a specific recommendation. Section 7 runs a full-blown Monte Carlo simulation exercise that demonstrates the strong performance of quadratic shrinkage in a wide variety of scenarios, matching the best RMT-based estimators that allow infinite degrees of nonlinearity. Section 8 concludes. An Appendix contains programming code, proofs of all mathematical results, and additional material such as supplementary Monte Carlo simulations.

2 Finite-Sample Analysis

Even though the sample size n is fixed in this section, we nonetheless subscript quantities by n to harmonize the notation throughout the paper.

2.1 General Setup

Let Σ_n denote a p-dimensional positive-definite population covariance matrix, where $2 < p < n$.[1] A mean-zero independent and identically distributed (i.i.d.) sample of n observations with covariance matrix Σ_n is arranged in an $n \times p$ matrix Y_n, which generates the sample covariance matrix $S_n := Y_n'Y_n/n$.[2] Its spectral decomposition is $S_n = U_n\Lambda_n U_n'$, where Λ_n is the diagonal matrix whose elements are the eigenvalues $\boldsymbol{\lambda}_n := (\lambda_{n,1}, \ldots, \lambda_{n,p})$ sorted in nondecreasing order without loss of generality, and an orthogonal matrix U_n whose columns $[u_{n,1} \ldots u_{n,p}]$ are the corresponding eigenvectors. Thus, $S_n = \sum_{i=1}^{p} \lambda_{n,i} \cdot u_{n,i}u_{n,i}'$.

2.2 Class of Estimators

Following Stein (1986, Lecture 4), we seek an estimator of the form $\widetilde{\Sigma}_n := U_n\widetilde{\Delta}_n U_n'$, where $\widetilde{\Delta}_n$ is a diagonal matrix whose elements $\widetilde{\boldsymbol{\delta}}_n := \left(\widetilde{\delta}_{n,1}, \ldots, \widetilde{\delta}_{n,p}\right) \in (0, +\infty)^p$ are a function of $\boldsymbol{\lambda}_n$. Such estimators are *rotation* equivariant because post-multiplying the data Y_n by an orthogonal matrix (with determinant one) rotates the estimators accordingly. By contrast, estimators from the sparsity literature such as Bickel and Levina (2008a,b) and El Karoui (2008) are merely *permutation* equivariant. This means that they are dependent on *a priori* information about the orientation of the orthonormal basis of the population eigenvectors which is impossible to verify in practice.

2.3 Loss Function

Any generic estimator $\widetilde{\Sigma}_n$ is evaluated according to the following loss function used by Stein (1975, 1986) and commonly referred to as Stein's loss:

Definition 2.1 (Stein's Loss)**.** *Let* $\mathsf{Tr}(\cdot)$ *denote the trace. Stein's loss is defined as:*

$$\mathcal{L}_n^{ST}\left(\Sigma_n, \widetilde{\Sigma}_n\right) := \frac{1}{p}\mathsf{Tr}\left(\Sigma_n^{-1}\widetilde{\Sigma}_n\right) - \frac{1}{p}\log\,\det\left(\Sigma_n^{-1}\widetilde{\Sigma}_n\right) - 1\ . \tag{2.1}$$

Proposition 2.1. *The solution to the optimization problem*

$$\underset{\widetilde{\Delta}_n\ \textit{diagonal}}{\operatorname{argmin}}\ \mathcal{L}_n^{ST}\left(\Sigma_n, U_n\widetilde{\Delta}_n U_n'\right)$$

$$\textit{is}\quad \overline{D}_n^{ST} := \mathsf{Diag}\left(\overline{d}_{n,1}^{ST}, \ldots, \overline{d}_{n,p}^{ST}\right)\quad \textit{where}\quad \overline{d}_{n,i}^{ST} := \frac{1}{u_{n,i}'\Sigma_n^{-1}u_{n,i}}\quad \textit{for } i = 1, \ldots, p. \tag{2.2}$$

[1] The singular case $p > n$ is covered in Section 5 and Appendix E.

[2] If the variables have a nonzero mean, the theorems in this paper also apply after in-sample demeaning, and adjusting the 'effective sample size' to $n - 1$ to account for the loss of one degree of freedom.

This results in an estimator, $\overline{S}_n^{\mathrm{ST}} := U_n \overline{D}_n^{\mathrm{ST}} U_n'$, which is not achievable in practice because Σ_n is unobservable, but constitutes a useful benchmark. At the qualitative level, we can already point out that the shrinkage formula (2.2) is in some sense 'the inverse of an inverse'. A less roundabout, more direct approach will be provided in Section 4.

2.4 *Bona Fide* Shrinkage Formula

Rewriting Equations (22)–(23) of Stein (1986, Lecture 4) in our notation, Stein approximates the unobservable $\overline{d}_{n,i}^{\mathrm{ST}}$'s with the *bona fide* estimator

$$\widetilde{d}_{n,i} := \frac{n\lambda_{n,i}}{n + p - 2i + 1 + 2\sum_{j=i+1}^{p} \frac{n\lambda_{n,j}}{n\lambda_{n,i} - n\lambda_{n,j}} - 2\sum_{j=1}^{i-1} \frac{n\lambda_{n,i}}{n\lambda_{n,j} - n\lambda_{n,i}}}, \tag{2.3}$$

for all $i = 1, \ldots, p$. The main differences between Stein's notation and ours are that the roles of the indices i and j are swapped, and his eigenvalues ℓ_j $(j = 1, \ldots, p)$ are equal to n times those of the sample covariance matrix. The resulting covariance matrix estimator is $\widetilde{S}_n := \sum_{i=1}^{p} \widetilde{d}_{n,i} \cdot u_{n,i} u_{n,i}'$. Although expression (2.6) below is the more common one in the literature, the original expression in Stein (1986, Lecture 4) is indeed (2.3).

Stein's estimator broke new ground and fathered an extensive literature on rotation-equivariant shrinkage estimation of a covariance matrix; for example, see the articles by Efron and Morris (1976), Haff (1980, 1991), Lin and Perlman (1985), Dey and Srinivasan (1985), Krishnamoorthy and Gupta (1989), Loh (1991), Pal (1993), Yang and Berger (1994), Daniels and Kass (2001), Ledoit and Wolf (2004, 2012), Chen et al. (2009), Won et al. (2013), Donoho et al. (2018), and the references therein.

2.5 The Stein Shrinker

One glaring issue with Stein's formula is that it is not intuitive because it is highly nonlinear. Therefore, our first contribution is to reinterpret it as linear shrinkage — but not of the sample eigenvalues: of their *inverses* instead. There is no *a priori* reason why linearly shrinking the inverse eigenvalues should be better than shrinking the eigenvalues themselves; it is just what Stein's mathematical discoveries lead to. But linearly shrinking the inverse eigenvalues is certainly no more worrisome than linearly shrinking the eigenvalues, an operation that has been well understood and accepted by researchers at least since Ledoit and Wolf (2004).

Theorem 2.1. *The nonlinear shrinkage formula* (2.3) *is mathematically equivalent to*

$$\forall i = 1, \ldots, p \qquad \widetilde{d}_{n,i}^{-1} = \left(1 - \frac{p-1}{n}\right) \lambda_{n,i}^{-1} + \left(\frac{p-1}{n}\right) \times 2\lambda_{n,i}^{-1} \widetilde{\theta}_n\left(\lambda_{n,i}^{-1}\right) \tag{2.4}$$

$$\textit{where} \quad \forall x \in \mathbb{R} \quad \widetilde{\theta}_n(x) := \frac{1}{p-1} \sum_{\substack{j=1 \\ \lambda_{n,j}^{-1} \neq x}}^{p} \lambda_{n,j}^{-1} \frac{1}{\lambda_{n,j}^{-1} - x}\,. \tag{2.5}$$

Proof of Theorem 2.1. From Equation (2.3) we deduce

$$\widetilde{d}_{n,i} = \frac{\lambda_{n,i}}{1 + \frac{p-1}{n} + \frac{2}{n}\sum_{j \neq i} \frac{\lambda_{n,j}}{\lambda_{n,i} - \lambda_{n,j}}} \tag{2.6}$$

$$\widetilde{d}_{n,i}^{-1} = \lambda_{n,i}^{-1} + \frac{p-1}{n}\lambda_{n,i}^{-1}\left[1 + 2\frac{1}{p-1}\sum_{j \neq i}\left(\frac{\lambda_{n,j}^{-1}}{\lambda_{n,j}^{-1} - \lambda_{n,i}^{-1}} - 1\right)\right] \tag{2.7}$$

$$= \left(1 - \frac{p-1}{n}\right)\lambda_{n,i}^{-1} + \left(\frac{p-1}{n}\right) \times \frac{1}{p-1}\sum_{j \neq i}\frac{2\lambda_{n,j}^{-1}\lambda_{n,i}^{-1}}{\lambda_{n,j}^{-1} - \lambda_{n,i}^{-1}} \ . \ \blacksquare \tag{2.8}$$

We call this shrinkage "linear" in inverse-eigenvalues space because it is a convex linear combination of $\lambda_{n,i}^{-1}$ with a shrinkage target that has a common structure independent of the shrinkage intensity. The fact that the shrinkage intensity is $(p-1)/n$ makes intuitive sense because more shrinkage must be applied when the curse of dimensionality is strong. The shrinkage target is a multiplicative modulation of the eigenvalue $\lambda_{n,i}^{-1}$ that is being shrunk. From now on we shall call the modulator

$$\widetilde{\theta}_n(x) = \frac{1}{p-1}\sum_{\substack{j=1 \\ \lambda_{n,j}^{-1} \neq x}}^{p} \lambda_{n,j}^{-1}\frac{1}{\lambda_{n,j}^{-1} - x} \tag{2.9}$$

the "Stein shrinker". It is locally adaptive in the sense that $\widetilde{\theta}_n(x)$ is not constant. This property stands in sharp contrast with the linear shrinkage formula of Ledoit and Wolf (2004) that shrinks all sample eigenvalues linearly towards the same common (global) target: their grand mean. Letting shrinkage targets adapt to local conditions makes it possible to extract additional accuracy gains over and above those already attained by Ledoit and Wolf (2004), especially when the sample eigenvalues are dispersed or clustered.

Visual inspection of the Stein shrinker immediately reveals that: 1) it attracts eigenvalues towards each other; 2) larger precision matrix eigenvalues have proportionally stronger power of attraction; 3) the intensity of the attraction vanishes to zero as the distance between eigenvalues increases; and 4) bad things happen (explosive numerical behavior) when two eigenvalues get too close to each other. The first three properties are features, but the fourth one can be considered a 'bug', and the next section is devoted to fixing it by smoothing.[3]

3 Optimal Linear-Inverse Shrinkage

Stein's analysis was only suggestive of optimality, not conclusive. First, he conceded that he had to ignore the effect of a certain derivatives term (Stein, 1986, p. 1391). Second, he post-processed the shrunken eigenvalues of Equation (2.3) through a numerical procedure called "isotonization" in order to restore the ordering of the eigenvalues, and ensure they are all positive. Rajaratnam

[3] Rajaratnam and Vincenzi (2016a,b) provide an in-depth study of the limitations of the original (or raw) Stein shrinker.

and Vincenzi (2016a) show that isotonization is actually essential to the empirical success of Stein's estimator, but its theoretical properties are extremely hard to investigate formally. Our next task is, therefore, to develop a *provably* optimal version of Stein's estimator that is purely analytical in nature. Given the lack of tractability in finite samples, we move to the framework of *large-dimensional asymptotics*.

3.1 Large-Dimensional Asymptotics

The idea is that the matrix dimension p and sample size n go to infinity together, while their ratio p/n (called the "concentration (ratio)") converges to some limit $c \in (0,1)$.[4] This framework is empirically relevant as soon as the matrix dimension is non-negligible with respect to the sample size. The following assumptions, or variations thereof, have been employed before in this literature, also known in physics as random matrix theory (RMT), going back to Wigner (1955).

Assumption 3.1 (Dimension)**.** *Let n denote the sample size and $p := p(n)$ the number of variables. It is assumed that the concentration (ratio) $c_n := p/n$ converges, as $n \to \infty$, to a limit $c \in (0,1)$ called the "limiting concentration (ratio)". Furthermore, there exists a compact interval included in $(0,1)$ that contains p/n for all n large enough.*

The elegant way to handle the ever-increasing dimension of the vector of eigenvalues is to map it into a function:

Definition 3.1. *The empirical distribution function (e.d.f.) of a collection of eigenvalues $(\alpha_1, \ldots, \alpha_p)$ is the nondecreasing step function $x \longmapsto p^{-1} \sum_{i=1}^{p} \mathbf{1}_{\{\alpha_i \leq x\}}$, where $\mathbf{1}$ denotes the indicator function.*

This e.d.f. returns the proportion of eigenvalues that lie weakly below its argument.

Assumption 3.2.

a. The population covariance matrix Σ_n is a nonrandom symmetric positive-definite matrix of dimension $p \times p$.

b. X_n is an $n \times p$ matrix of i.i.d. random variables with mean zero, variance one, and finite 16th moment. The matrix of observations is $Y_n := X_n \times \sqrt{\Sigma_n}$. Neither $\sqrt{\Sigma_n}$ nor X_n are observed on their own: only Y_n is observed.

c. Let $\boldsymbol{\tau}_n := (\tau_{n,1}, \ldots, \tau_{n,p})'$ denote a system of eigenvalues of Σ_n, and H_n the e.d.f. of population eigenvalues. It is assumed that H_n converges weakly to a limit law H, called the "limiting spectral distribution (function)".

d. $\mathsf{Supp}(H)$, the support of H, is the union of a finite number of closed intervals, bounded away from zero and infinity. Also, there exists a compact interval $[\underline{T}, \overline{T}] \subset (0, \infty)$ that contains $\{\tau_{n,1}, \ldots, \tau_{n,p}\}$ for all n large enough.

[4] The case $c \in (1, +\infty)$ is covered in Section 5 and Appendix E.

Remark 3.1. The assumption of a finite 16th moment comes from Theorem 3 of Jing et al. (2010), which we use in our proofs. However, these authors' Remark 1 conjectures that a finite fourth moment is actually enough, and our own Monte Carlo simulations in Section 7 concur. ■

The literature on the eigenvalues of the sample covariance matrix under large-dimensional asymptotics is based on a foundational result by Marčenko and Pastur (1967), which has been strengthened and broadened by subsequent authors including Silverstein and Bai (1995) and Silverstein (1995), among others. The latter's Theorem 1.1 implies that, under Assumptions 3.1–3.2, there exists a continuous non-stochastic limiting sample spectral distribution function F such that the e.d.f. of the sample eigenvalues, denoted by F_n, converges pointwise almost surely to F. This limiting sample spectral c.d.f. F is uniquely determined by c and H, so we will denote it more explicitly by $F_{c,H}$ whenever there is some risk of ambiguity. Assumption 3.2 together with Theorem 1.1. of Bai and Silverstein (1998) imply that the support of F, denoted by $\mathsf{Supp}(F)$, is the union of a finite number $\nu \geq 1$ of compact intervals: $\mathsf{Supp}(F) = \bigcup_{k=1}^{\nu}[a_k, b_k]$, where $0 < a_1 < b_1 < \cdots < a_\nu < b_\nu < \infty$.

Following Ledoit and Wolf (2018), we extend the class of rotation-equivariant covariance matrix estimators from Section 2 into the realm of large-dimensional asymptotics.

Definition 3.2 (Class of Estimators)**.** *Covariance matrix estimators are of the type* $\widetilde{\Sigma}_n := U_n \widetilde{\Delta}_n U_n'$*, where* $\widetilde{\Delta}_n$ *is a diagonal matrix:* $\widetilde{\Delta}_n := \mathsf{Diag}\left(\widetilde{\delta}_n(\lambda_{n,1}) \ldots, \widetilde{\delta}_n(\lambda_{n,p})\right)$*, and* $\widetilde{\delta}_n$ *is a (possibly random) real univariate function which can depend on* S_n*.*

Every candidate shrinkage function $\widetilde{\delta}_n$ must behave well asymptotically:

Assumption 3.3 (Limiting Shrinkage Function)**.** *There exists a nonrandom real univariate function* $\widetilde{\delta}$ *defined on* $\mathsf{Supp}(F)$ *and continuously differentiable such that* $\widetilde{\delta}_n(x) \xrightarrow{\text{a.s.}} \widetilde{\delta}(x)$*, for all* $x \in \mathsf{Supp}(F)$*. Furthermore, this convergence is uniform over* $x \in \bigcup_{k=1}^{\nu}[a_k + \eta, b_k - \eta]$*, for any small* $\eta > 0$*. Finally, for any small* $\eta > 0$*, there exists a finite nonrandom constant* $\widetilde{K}$ *such that almost surely, over the set* $x \in \bigcup_{k=1}^{\nu}[a_k - \eta, b_k + \eta]$*,* $\widetilde{\delta}_n(x)$ *is uniformly bounded by* $\widetilde{K}$ *from above and by* $1/\widetilde{K}$ *from below, for* n *large enough.*

3.2 Smoothed Stein Shrinker

Our second contribution is to prove that a simple smoothing of the Stein shrinker yields an optimal estimator under large-dimensional asymptotics, even without requiring the variates to be normally distributed.

Theorem 3.1. *Suppose Assumptions 3.1–3.3 hold. Then, for any covariance matrix estimator* $\widetilde{\Sigma}_n$ *in the rotation-equivariant class of Definition 3.2, Stein's loss* $\mathcal{L}_n^{ST}(\Sigma_n, \widetilde{\Sigma}_n)$ *converges almost surely to a nonrandom limit as* p *and* n *go to infinity together. This limit is minimized if*

$\widetilde{\delta}_n(\lambda_{n,i}) = \widehat{d}_{n,i}$, *with* $\widehat{d}_{n,i}$ *satisfying:*

$$\forall i = 1, \ldots, p \qquad \widehat{d}_{n,i}^{-1} := \left(1 - \frac{p}{n}\right) \lambda_{n,i}^{-1} + \left(\frac{p}{n}\right) \times 2\lambda_{n,i}^{-1} \widehat{\theta}_n\left(\lambda_{n,i}^{-1}\right) \tag{3.1}$$

$$\textit{where} \quad \forall x \in \mathbb{R} \qquad \widehat{\theta}_n(x) := \frac{1}{p} \sum_{j=1}^{p} \lambda_{n,j}^{-1} \frac{\lambda_{n,j}^{-1} - x}{\left(\lambda_{n,j}^{-1} - x\right)^2 + h_n^2 \lambda_{n,j}^{-2}} \ , \tag{3.2}$$

$$\textit{and a smoothing parameter} \quad h_n \sim K n^{-\alpha} \quad \textit{for some } K > 0 \textit{ and } \alpha \in (0, 2/5) \ . \tag{3.3}$$

The resulting covariance matrix estimator is $\widehat{S}_n := \sum_{i=1}^{p} \widehat{d}_{n,i} \cdot u_{n,i} u_{n,i}'$.

Proofs are in the appendix. As a technical aside, the proofs in this paper build upon Jing et al. (2010) and Ledoit and Wolf (2020). We use the most salient elements of both works as stepping stones to make further headway. From Jing et al. (2010) we borrow techniques that enable us to extend the analysis of Ledoit and Wolf (2020) from kernels with bounded support to kernels with unbounded support. From Ledoit and Wolf (2020), we borrow techniques that enable us to extend the kernel estimation of the limiting sample spectral density in Jing et al. (2010) to its Hilbert transform. Beyond both papers, we move into the realm of the *inverses* of the sample eigenvalues (as opposed to the sample eigenvalues themselves), and of the first incomplete moment function.

Equation (3.1) is still linear shrinkage of the inverse sample eigenvalues. Replacing the shrinkage intensity $(p-1)/n$ from Equation (2.4) with p/n is immaterial, as we operate under large-dimensional asymptotics. What matters is that, inside the summation, the discontinuous, explosive influence function

$$\frac{1}{\lambda_{n,j}^{-1} - x} \quad \text{is replaced by a smoother equivalent} \quad \frac{\lambda_{n,j}^{-1} - x}{\left(\lambda_{n,j}^{-1} - x\right)^2 + h_n^2 \lambda_{n,j}^{-2}} \ . \tag{3.4}$$

We view the covariance matrix estimator $\widehat{S}_n$ of Theorem 3.1 as linear shrinkage in inverse-eigenvalues space, or linear-inverse shrinkage (LIS) for short. It can also be interpreted as linear shrinkage of the precision matrix.

The bandwidth parameter h_n controls the degree of smoothing. If h_n were equal to zero, the two fractions in Equation (3.4) would be mathematically identical. For the purpose of the proofs, we require h_n to be strictly positive but to vanish asymptotically as n goes to infinity. In terms of nomenclature, we call $\widehat{\theta}_n(x)$ the "smoothed Stein shrinker". Figure 1 illustrates visually.

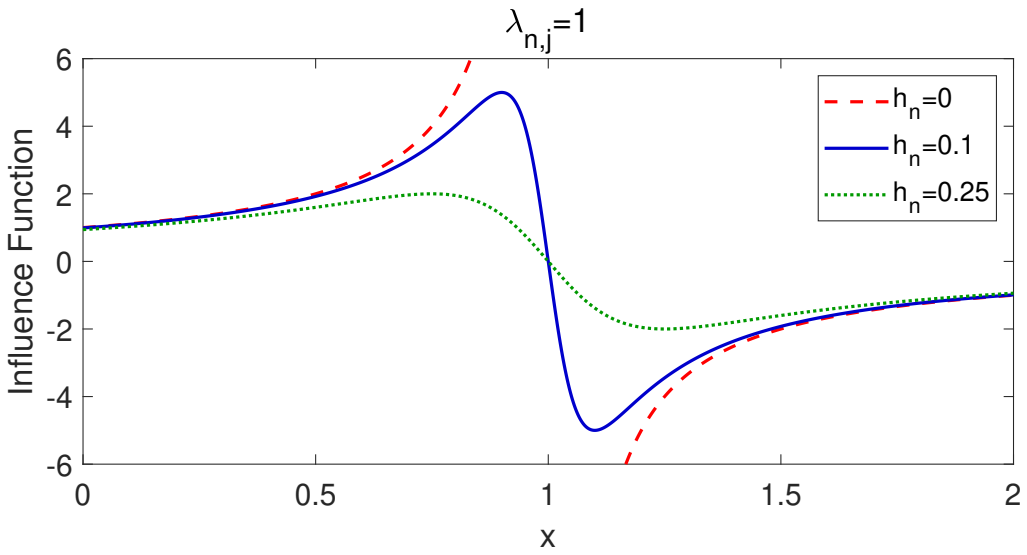

Figure 1: Dependence of the influence function on the regularization parameter h_n. When $h_n = 0$, the function diverges, which generates much numerical instability.

Stein's formula required *ex post* numerical regularization through the *ad hoc* procedure of isotonization. We avoid this problem by bringing regularization analytically inside the formula through the introduction of the parameter h_n into the smoothed shrinkage modulator $\widehat{\theta}_n(x)$. Not only does Theorem 3.1 formally prove optimality, but it does so without requiring normality, as can be seen from Assumption 3.2(b).

Careful examination of the influence function shows a phenomenon of "local shrinkage". If $\lambda_{n,i}^{-1}$ is slightly below $\lambda_{n,j}^{-1}$ ($x < 1$ in Figure 1), then the influence exerted by $\lambda_{n,j}^{-1}$ onto $\lambda_{n,i}^{-1}$ is positive, meaning that $\widehat{d}_{n,i}^{-1}$ will tend to go up *towards* $\lambda_{n,j}^{-1}$ (everything else being equal). Similarly, if $\lambda_{n,i}^{-1}$ is slightly above $\lambda_{n,j}^{-1}$ ($x > 1$ in Figure 1), then the influence exerted by $\lambda_{n,j}^{-1}$ onto $\lambda_{n,i}^{-1}$ is negative, meaning that $\widehat{d}_{n,i}^{-1}$ will tend to go down, *also* towards $\lambda_{n,j}^{-1}$. Thus, there is shrinkage in the sense that inverse eigenvalues tend to be attracted towards one another. But, unlike the linear shrinkage formula of Ledoit and Wolf (2004), this shrinkage is "local" because the influence exerted by distant eigenvalues vanishes quickly. This is why this particular form of linear shrinkage can generate substantial improvements when the population eigenvalues are dispersed, clustered, or otherwise unruly.

At this stage, there is nothing guaranteeing that the shrunken inverse eigenvalues will be strictly positive. This was also the case in the original Stein estimator, which is part of the reason why he had recourse to isotonization — which fixed the problem. But since we would prefer to retain a purely analytic formula, we propose a minor alteration to Equation (3.1). It is based on the observation that shrinkage always operated inwards at the extremities of the

support. Thus, in particular, all $\widehat{d}_{n,i}^{-1}$ should be greater than or equal to $\min_{j=1,\ldots,p}\left(\lambda_{n,j}^{-1}\right) = \lambda_{n,p}^{-1}$, which is itself strictly positive. Hence the correction:

$$\left(\widehat{d}_{n,i}^{\text{LIS}}\right)^{-1} := \max\left[\lambda_{n,p}^{-1}, \frac{1}{p}\sum_{j=1}^{p}\lambda_{n,j}^{-1}\frac{\lambda_{n,j}^{-1}-x}{\left(\lambda_{n,j}^{-1}-x\right)^2+h_n^2\lambda_{n,j}^{-2}}\right] \text{ and } \widehat{S}_n^{\text{LIS}} := \sum_{i=1}^{p}\widehat{d}_{n,i}^{\text{LIS}}\cdot u_{n,i}u_{n,i}' \ . \quad (3.5)$$

This modification is not needed asymptotically, so the constrained estimator shares the optimality properties stated in Theorem 3.1. Yet, we prefer it because it is safer in finite samples.[5]

4 Shrinkage Under Alternative Loss Functions

Stein (1986, p. 1390) candidly acknowledges that he uses the loss function of Definition 2.1, now called Stein's loss, "because it is comparatively easy to work with this loss function." The technological boost from large-dimensional asymptotics now gives us a chance to solve the same problem for loss functions that are "comparatively hard to work with", but potentially more interesting for a wide variety of practical applications.

4.1 Inverse Stein's Loss

To facilitate continuity with Section 3, we start by applying Stein's loss to the precision matrix, as Tsukuma (2005) did.

Definition 4.1 (Inverse Stein's Loss)**.** *The Inverse Stein's loss is defined as:*

$$\mathcal{L}_n^{IS}\left(\Sigma_n,\widetilde{\Sigma}_n\right) := \mathcal{L}_n^{ST}\left(\widetilde{\Sigma}_n,\Sigma_n\right) = \mathcal{L}_n^{ST}\left(\Sigma_n^{-1},\widetilde{\Sigma}_n^{-1}\right) = \frac{1}{p}\mathsf{Tr}\left(\Sigma_n\widetilde{\Sigma}_n^{-1}\right) - \frac{1}{p}\log\det\left(\Sigma_n\widetilde{\Sigma}_n^{-1}\right) - 1 \ . \quad (4.1)$$

In finite samples, the Frobenius loss, the Minimum Variance loss and the Inverse Stein's loss all give rise to the same optimal oracle estimator defined by Equation (4.4).

Proposition 4.1. *The solution to the optimization problem*

$$\underset{\widetilde{\Delta}_n \ \textit{diagonal}}{\text{argmin}} \ \mathcal{L}_n^{IS}\left(\Sigma_n, U_n\widetilde{\Delta}_nU_n'\right)$$

$$\textit{is} \quad \overline{D}_n := \mathsf{Diag}\left(\overline{d}_{n,1},\ldots,\overline{d}_{n,p}\right) \quad \textit{where} \quad \overline{d}_{n,i} := u_{n,i}'\Sigma_n u_{n,i} \quad \textit{for } i=1,\ldots,p \ . \quad (4.2)$$

This results in an estimator, $\overline{S}_n := U_n\overline{D}_nU_n'$, which is not achievable in practice because Σ_n is unobservable, but constitutes a useful benchmark. For this reason, it is called an "oracle" estimator. Comparing with Stein's loss, in terms of covariance matrix eigenvalues we move from $\left(u_{n,i}'\Sigma_n^{-1}u_{n,i}\right)^{-1}$ to $u_{n,i}'\Sigma_n u_{n,i}$. This move looks like rather intuitive and easily understandable.

[5] We thank two referees for having prompted us to find a way to guarantee the strict positivity of shrunken inverse eigenvalues.

4.2 Frobenius Loss

Definition 4.2 (Frobenius Loss)**.** *The Frobenius loss is defined as:*

$$\mathcal{L}^{FR}(\Sigma_n, \widetilde{\Sigma}_n) := \frac{1}{p}\mathsf{Tr}\left[\left(\Sigma_n - \widetilde{\Sigma}_n\right)^2\right] \ . \tag{4.3}$$

The linear shrinkage formula of Ledoit and Wolf (2004) has popularized Frobenius-loss-based covariance matrix estimation in fields as far apart as cancer research (Pyeon et al., 2007), macroeconomics (Korniotis, 2008), brain-computer interface (Vidaurre et al., 2009), psychology (Markon, 2010), political science (Tenenhaus and Tenenhaus, 2011), analytical chemistry (Guo et al., 2012), geology (Elsheikh et al., 2013), and neuroscience (Deligianni et al., 2014), to highlight but a selected few.

This literature also incorporates numerous extensions, adaptations and refinements of Ledoit and Wolf's linear shrinkage estimator. For example, Schäfer and Strimmer (2005) propose six different shrinkage targets. Stoica et al. (2008) embed linear shrinkage into space-time adaptive processing, an important radar technique. Chen et al. (2011) improve the formula for the shrinkage intensity when variates are normally, respectively elliptically, distributed. What such papers have in common is that they all set out to minimize the expected Frobenius loss.

Proposition 4.2. *The solution to the optimization problem*

$$\underset{\widetilde{\Delta}_n \ \textit{diagonal}}{\text{argmin}} \quad \mathcal{L}_n^{FR}\left(\Sigma_n, U_n\widetilde{\Delta}_n U_n'\right)$$

$$\textit{is} \quad \overline{D}_n := \mathsf{Diag}\left(\overline{d}_{n,1}, \ldots, \overline{d}_{n,p}\right) \quad \textit{where} \quad \overline{d}_{n,i} := u_{n,i}'\Sigma_n u_{n,i} \quad \textit{for } i = 1, \ldots, p \ . \tag{4.4}$$

4.3 Minimum Variance Loss

An even stronger justification for moving away from Stein's loss comes from reviewing typical applications that make use of the covariance matrix and its inverse, in order to craft a tailor-made loss function that directly speaks to their overarching objectives. Markowitz (1952) essentially launched finance as a scientific field. His key contribution was to show that the investor should think in terms of a portfolio allocated across a multitude of candidate financial assets, and trade off expected returns (good) against total portfolio risk (bad). In this context, the variables whose covariances we estimate are asset returns (potentially in excess of the risk-free rate). Given a $p \times 1$ vector of expected (excess) returns μ, the optimal trade-off is achieved by solving the minimization problem

$$\underset{w \in \mathbb{R}^p}{\text{argmin}}\, w'\Sigma_n w \qquad \text{subject to: } w'\mu = \gamma \ , \tag{4.5}$$

where γ is an expected return target. This is also known as the "tangency portfolio". The solution to (4.5) is of the form $w = \text{scalar} \times \Sigma_n^{-1}\mu$, where the scalar multiplier is chosen to satisfy the investor's capital and leverage constraints. In reality, the population covariance matrix Σ_n is unobservable, so we must replace it with some estimator $\widetilde{\Sigma}_n$, and the question is how high is

the *out-of-sample* volatility $\widetilde{w}'\Sigma_n\widetilde{w}$ of the *in-sample efficient* portfolio $\widetilde{w} := \widetilde{\Sigma}_n^{-1}\mu$. Given that every investor has a different μ based on individual expectations of how assets will perform, it is desirable to 'average out', or somehow 'integrate' the answer across all possible directions of the expected return vector μ.

The above framework carries through one-for-one as we move to Capon (1969) beamforming in signal processing (radar, sonar, wireless communications, seismology, etc). When a narrow-band source impinging upon an array of sensors, the vector μ should be interpreted as a response vector that includes effects such as coupling between elements and subsequent amplification, presumed to be known due to the design of the sensor array, called $\boldsymbol{a}(\boldsymbol{\theta}_s)$ by Abrahamsson et al. (2007) in their Equation (2). The rest of the analysis is identical to the finance application.

It is the same for optimal fingerprinting, a method originally due to Hasselmann (1993) that has been chosen by the Intergovernment Panel on Climate Change to track global warming (IPCC, 2007, Section 9.A.1). In this context, the place of the vector μ is taken by the nonrandom response of the earth-system to an external forcing (Ribes et al., 2009, Section 2.1). The objective is to find a $p \times 1$ "fingerprint" vector so that the linear combination of temperature measurements weighted by the entries of the fingerprint minimizes climate-variability noise subject to a linear constraint on signal intensity.

Yet another method that fits in this framework is linear discriminant analysis (LDA), an essential tool for machine learning. In the two-class case, the vector μ represents the difference between the average score of one class across all dimensions of measurement, minus the average score of the other class against which we wish to discriminate. The objective of LDA is to find a one-dimensional subspace in which the classes are well separated. This is achieved by requiring that, after projection onto the subspace, the ratio of between-class variance to within-class variance is maximal. In this context, $\widetilde{w}$ is the direction of the one-dimensional subspace used to discriminate between classes. LDA has been used extensively in efforts to develop an interface between the brain and a computer (Vidaurre et al., 2009).

Motivated by this seemingly ubiquitous mathematical problem, Engle et al. (2019) advocate what is called the "Minimum Variance" loss function:

Definition 4.3 (Minimum Variance Loss)**.** *The Minimum Variance loss is defined as:*

$$\mathcal{L}^{MV}\big(\Sigma_n, \widetilde{\Sigma}_n\big) := \frac{\mathsf{Tr}\big(\widetilde{\Sigma}_n^{-1}\Sigma_n\widetilde{\Sigma}_n^{-1}\big)/p}{\Big[\mathsf{Tr}\big(\widetilde{\Sigma}_n^{-1}\big)/p\Big]^2} - \frac{1}{\mathsf{Tr}\big[\Sigma_n^{-1}\big]/p} \ . \tag{4.6}$$

It represents the *true* variance of a linear combination of the original p variables selected to have minimum *estimated* variance subject to a generic linear constraint, suitably normalized under large-dimensional asymptotics. Thus, it is extremely relevant to the empirical applications enumerated above and mathematically similar ones in other fields of science. We will not go more in-depth here into the justification of this particular loss function because it has been given already in Engle et al. (2019, Section 4), as well as in a precursor paper by Engle and Colacito (2006, Section 2). The good news is that we do not have to choose between Inverse Stein, Frobenius, and Minimum Variance loss because they all lead to the same oracle estimator.

Proposition 4.3. *The oracle estimator* $\overline{S}_n := \sum_{i=1}^{p} \overline{d}_{n,i} \cdot u_{n,i} u'_{n,i}$ *minimizes the loss function* $\mathcal{L}_n^{MV}$ *within the class of rotation-equivariant estimators specified in Section 2.2.*

The profound agreement between three seemingly different loss functions: $\mathcal{L}^{\text{MV}}$, $\mathcal{L}^{\text{FR}}$ and $\mathcal{L}_n^{IS}$, serves as further justification for adopting this family.

Remark 4.1. The common feature between all three of the loss functions in this family is that they are based on the population covariance matrix Σ_n, and not on the inverse Σ_n^{-1} used by Stein's loss. This is desirable because, if one of the population eigenvalues happens to be very close to zero, which can be extremely hard to detect when $p > n$ as in Section 5, inverting Σ_n would generate much numerical instability. ■

4.4 Conjugate and Amplitude of the Smoothed Stein Shrinker

To construct a *bona fide* estimator, that is, one that only depends on the observable data Y_n, we go back to the more tractable framework of large-dimensional asymptotics. To express our solution, we need to present two closely related concepts borrowed from signal processing: the *conjugate* and the *amplitude*. As a sneak preview, the new loss functions will introduce a second shrinkage target governed by the squared amplitude of the smoothed Stein shrinker.

The notion of conjugate goes all the way back to the *analytic signal* theory of Gabor (1946). The basic idea is that what we observe (which, in our case, is the smoothed Stein shrinker $\widehat{\theta}_n(x)$) also encrypts a conjugate that is not directly observable but is extractable via the *Hilbert transform*. This important transform is defined as convolution with the Cauchy kernel $(\pi t)^{-1}$:

$$\forall x \in \mathbb{R} \qquad \widehat{\theta}_n^*(x) = \mathcal{H}_{\widehat{\theta}_n}(x) := \frac{1}{\pi} PV \int_{-\infty}^{+\infty} \widehat{\theta}_n(t) \frac{dt}{t - x} \ , \tag{4.7}$$

where PV stands for the *Cauchy principal value*, which is used to evaluate the singular integral:

$$PV \int_{-\infty}^{+\infty} \widehat{\theta}_n(t) \frac{dt}{t - x} := \lim_{\varepsilon \to 0^+} \left[\int_{-\infty}^{x-\varepsilon} \widehat{\theta}_n(t) \frac{dt}{t - x} + \int_{x+\varepsilon}^{+\infty} \widehat{\theta}_n(t) \frac{dt}{t - x} \right] \ . \tag{4.8}$$

Conjugation is anti-involutive, meaning that the conjugate of the conjugate is none other than the original function itself (up to a minus sign). For example, the conjugate of the sine is the cosine, and the conjugate of the cosine is minus the sine. So the two conjugates are best thought of as a pair 'joined at the hip'. In our case, it is worth finding out the interpretation of the conjugate of the Stein shrinker.

Proposition 4.4. *The smoothed Stein shrinker*

$$\widehat{\theta}_n(x) = \frac{1}{p} \sum_{j=1}^{p} \lambda_{n,j}^{-1} \frac{\lambda_{n,j}^{-1} - x}{\left(\lambda_{n,j}^{-1} - x\right)^2 + h_n^2 \lambda_{n,j}^{-2}} \tag{4.9}$$

has for its conjugate

$$\widehat{\theta}_n^*(x) = \frac{1}{p} \sum_{j=1}^{p} \lambda_{n,j}^{-1} \frac{h_n \lambda_{n,j}^{-1}}{\left(\lambda_{n,j}^{-1} - x\right)^2 + h_n^2 \lambda_{n,j}^{-2}} \ . \tag{4.10}$$

We see that (4.10) is a nonparametric estimator of the derivative of the first incomplete moment of the spectral distribution of the sample precision matrix, using the Cauchy density as kernel, and using the locally adaptive bandwidth $h_{n,j} := h_n \lambda_{n,j}^{-1}$. Thus, the conjugate $\widehat{\theta}_n^*(x)$ shows where the inverse eigenvalues lie, like a density that would overweight the larger inverse eigenvalues. It is intuitively satisfying that the attraction force between eigenvalues is the conjugate of their location, and vice-versa.

The analytic signal theory of Gabor (1946) goes one step further to define the *amplitude* of a signal by combining the original function with its conjugate in a quadratic way. Vakman (1996, Section II) proves Gabor's formula captures our physical intuition about what amplitude should mean. Formally, the squared amplitude is defined as follows:

$$\mathcal{A}^2_{\widehat{\theta}_n}(x) := \widehat{\theta}_n(x)^2 + \widehat{\theta}_n^*(x)^2 = \left[-\mathcal{H}_{\widehat{\theta}_n^*}(x)\right]^2 + \left[\mathcal{H}_{\widehat{\theta}_n}(x)\right]^2 \tag{4.11}$$

$$= \left[\frac{1}{p}\sum_{j=1}^{p} \lambda_{n,j}^{-1} \frac{\lambda_{n,j}^{-1} - x}{\left(\lambda_{n,j}^{-1} - x\right)^2 + h_n^2 \lambda_{n,j}^{-2}}\right]^2 + \left[\frac{1}{p}\sum_{j=1}^{p} \lambda_{n,j}^{-1} \frac{h_n \lambda_{n,j}^{-1}}{\left(\lambda_{n,j}^{-1} - x\right)^2 + h_n^2 \lambda_{n,j}^{-2}}\right]^2 . \tag{4.12}$$

The amplitude measures whether there is any action in terms of either attraction force (the shrinker $\widehat{\theta}_n$) or weighted density (its conjugate $\widehat{\theta}_n^*$); it puts both pair members on the same footing and 'envelopes' them. Near the outskirts, there is less action of either type, so the amplitude vanishes. Note the elegant symmetry in (4.12): The denominator of the fraction is the sum of the squares of two terms, and both terms take turns appearing in the numerators of $\widehat{\theta}_n(x)$ and $\widehat{\theta}_n^*(x)$, respectively. To illustrate, Figure 2 displays a shrinker that behaves almost like minus the cosine over the interval $[\pi, 2\pi]$, up to rescaling.

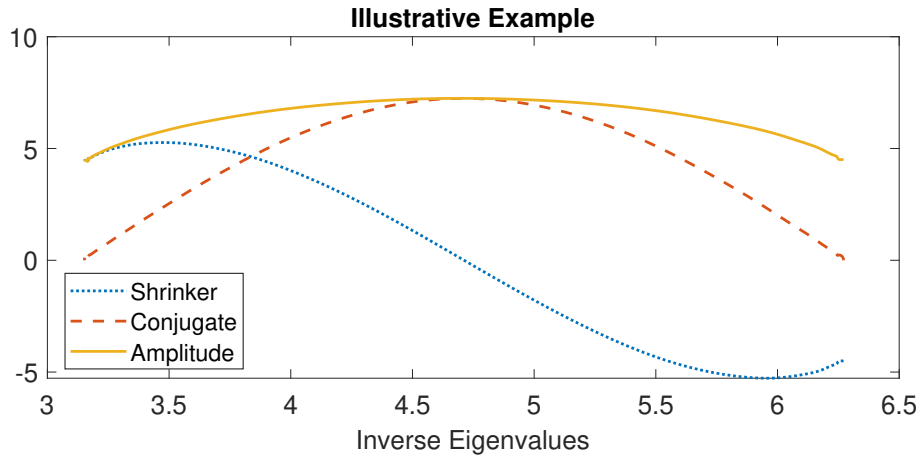

Figure 2: Stylized example of a shrinker, its conjugate and their amplitude.

As expected, its conjugate is basically minus the sine (up to rescaling), with a peak at the support midpoint of $3\pi/2$. The amplitude looks relatively flat because the cosine (attraction force) and the sine (weighted density) tend to compensate for each other. Near the edges of the support, the amplitude starts dipping because the action is dying down. To recapitulate:

- In the middle of the cluster of inverse eigenvalues, the amplitude is strong because the weighted density (the conjugate) is strongly positive.

- On either side of the cluster of inverse eigenvalues, the amplitude is still strong, but for a different reason: because the attraction force (the shrinker) is strong in absolute value.
- Further out near the edges of the support, the inverse eigenvalues are sparser and exert less pull, so the amplitude weakens.

4.5 Quadratic-Inverse Shrinkage (QIS) Estimator

We are now ready to state our third contribution, which is the adaptation of Stein's (smoothed) shrinkage formula to the Frobenius loss function and its two cousins:

Theorem 4.1. *Suppose Assumptions 3.1–3.3 hold. Then, for any covariance matrix estimator $\widehat{\Sigma}_n$ in the rotation-equivariant class of Definition 3.2, the Frobenius loss $\mathcal{L}_n^{FR}(\Sigma_n, \widehat{\Sigma}_n)$ converges in probability to a nonrandom limit as n goes to infinity. This limit is minimized if $\widetilde{\delta}_n(\lambda_{n,i}) = \widehat{\delta}_{n,i}$, with $\widehat{\delta}_{n,i}$ satisfying:*

$$\widehat{\delta}_{n,i}^{-1} = \left(1 - \frac{p}{n}\right)^2 \lambda_{n,i}^{-1} + 2\frac{p}{n}\left(1 - \frac{p}{n}\right) \lambda_{n,i}^{-1} \widehat{\theta}_n\left(\lambda_{n,i}^{-1}\right) + \left(\frac{p}{n}\right)^2 \lambda_{n,i}^{-1} \mathcal{A}_{\widehat{\theta}_n}^2\left(\lambda_{n,i}^{-1}\right), \text{ where} \tag{4.13}$$

$$\widehat{\theta}_n(x) := \frac{1}{p}\sum_{j=1}^{p} \lambda_{n,j}^{-1} \frac{\lambda_{n,j}^{-1} - x}{\left(\lambda_{n,j}^{-1} - x\right)^2 + h_n^2 \lambda_{n,j}^{-2}} \quad \text{and} \tag{4.14}$$

$$\mathcal{A}_{\widehat{\theta}_n}^2(x) = \left[\frac{1}{p}\sum_{j=1}^{p} \lambda_{n,j}^{-1} \frac{\lambda_{n,j}^{-1} - x}{\left(\lambda_{n,j}^{-1} - x\right)^2 + h_n^2 \lambda_{n,j}^{-2}}\right]^2 + \left[\frac{1}{p}\sum_{j=1}^{p} \lambda_{n,j}^{-1} \frac{h_n \lambda_{n,j}^{-1}}{\left(\lambda_{n,j}^{-1} - x\right)^2 + h_n^2 \lambda_{n,j}^{-2}}\right]^2, \tag{4.15}$$

for a smoothing parameter h_n satisfying the conditions of Equation (3.3).

Note that the first two terms on the right-hand side of Equation (4.13) are the same as the ones from Theorem 3.1, up to rescaling by a factor of $1 - (p/n)$. We call Equation (4.13) "quadratic shrinkage" of the inverse sample covariance matrix eigenvalues because the three weighting coefficients are quadratic functions of the concentration ratio p/n adding up to a perfect square:

$$\left(1 - \frac{p}{n}\right)^2 + 2\frac{p}{n}\left(1 - \frac{p}{n}\right) + \left(\frac{p}{n}\right)^2 = \left(1 - \frac{p}{n} + \frac{p}{n}\right)^2 = 1\,,$$

and also because the new term is the square of the amplitude of the smoothed Stein shrinker. Mathematically speaking, this new term goes back to the oracle formula for the Frobenius loss first proven in Ledoit and Péché (2011, Theorem 4); in particular, it is due to the squared modulus of the Stieltjes-transform term in their Equation (13).

Corollary 4.1. *The results stated in Theorem 4.1 also hold true for the Minimum Variance loss function $\mathcal{L}^{MV}$ and the Inverse Stein's loss function $\mathcal{L}^{IS}$, using the same shrinkage formula.*

The inverse sample covariance matrix eigenvalue is attracted to two shrinkage targets modulated, respectively, by $\widehat{\theta}_n(\lambda_{n,i}^{-1})$ and $\mathcal{A}_{\widehat{\theta}_n}^2(\lambda_{n,i}^{-1})$. The first target dominates mid-level concentration ratios, and the second one high-level concentration ratios. Figure 3 illustrates these shapes.

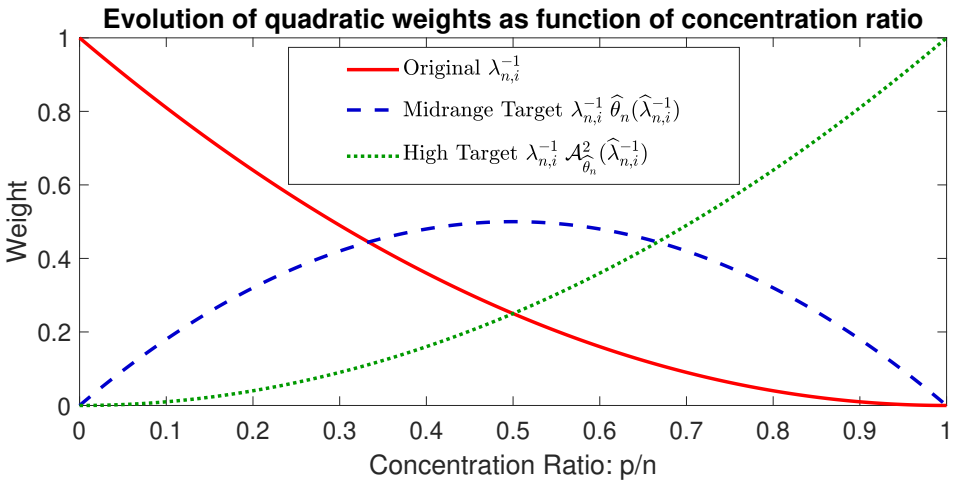

Figure 3: Evolution of the quadratic weights as a function of the concentration ratio p/n. The three of them sum up to one for every value of p/n.

Remark 4.2. The shrunken inverse eigenvalues $\{\widehat{\delta}_{n,i}^{-1}\}_{i=1}^{p}$ are guaranteed to be strictly positive, which was not necessarily the case when using Stein's loss in Section 2 and Theorem 3.1. Hence, a constraint analogous to Equation (3.5) is no longer required and shall not be applied here. As for undoing eigenvalue order violations through post-processing by a numerical algorithm, Appendix F.1 shows that it is not particularly useful here because the violations are relatively few, their magnitudes are benign, and it would not really move the needle in terms of accuracy of the QIS estimator. ■

Theorem 4.1 states that the quadratic shrinkage formula is optimal among *all* nonlinear shrinkage formulas; therefore, it cannot be beaten by any cubic or higher-order shrinkage. This result that reduces the most complicated nonlinear problem to a simple quadratic by tweaking Stein's classic formula is both powerful and mathematically elegant.

A potentially valuable refinement, more useful for the Frobenius and Inverse Stein losses than for Minimum Variance loss, is to rescale the quadratic-inverse shrinkage estimator to have the same trace as the sample covariance matrix, a property already enjoyed by the linear shrinkage of Ledoit and Wolf (2004):

$$\widehat{\Sigma}_n^{\text{QIS}} := \frac{\mathsf{Tr}\big[S_n\big]}{\mathsf{Tr}\big[\widehat{\Sigma}_n\big]}\,\widehat{\Sigma}_n \qquad \text{where} \qquad \widehat{\Sigma}_n := \sum_{i=1}^{p} \widehat{\delta}_{n,i} \cdot u_{n,i} u_{n,i}' \;. \tag{4.16}$$

This modification is not needed asymptotically, but may boost finite-sample performance in some applications. The bandwidth parameter h_n will be further specified in the next section.

Remark 4.3. For researchers who prefer a loss function that views eigenvalues close to zero as being 'as far away' as eigenvalues close to infinity, the Symmetrized Kullback-Leibler divergence $\mathcal{L}_n^{\text{SKL}}\big(\Sigma_n, \widetilde{\Sigma}_n\big) := \frac{1}{2p}\mathsf{Tr}\big(\Sigma_n^{-1}\widetilde{\Sigma}_n + \Sigma_n\widetilde{\Sigma}_n^{-1}\big) - 1$ of Moakher and Batchelor (2006, Eq. (17.8)) is invariant to matrix inversion. It is a convenient alternative to the affine-equivariant geodesic norm on the manifold of positive-definite matrices and to the Log-Euclidian norm. The two latter

norms are invariant to matrix inversion as well — see Förstner and Moonen (1999, Section 3.2) and Arsigny et al. (2006, p.413), respectively — but they are less tractable in the present framework. The rotation-equivariant covariance matrix estimator asymptotically optimal with respect to the Symmetrized Kullback-Leibler divergence is constructed by geometrically averaging Linear-Inverse shrinkage with Quadratic-Inverse shrinkage:

$$\widehat{S}^{\mathrm{SKL}} := \sum_{i=1}^{p} \sqrt{\widehat{S}^{\mathrm{LIS}} \times \widehat{\Sigma}^{\mathrm{QIS}}} \cdot u_{n,i} u_{n,i}' \ . \tag{4.17}$$

This result is a direct consequence of Theorems 3.1 and 4.1, conducted in the spirit of Ledoit and Wolf (2018, Section 4.4). For reasons detailed in the next section, this approach only works for $p < n$, but not in the singular case $p > n$. ■

Remark 4.4. It is worth noting that a loss function that pertains to the precision matrix such as Inverse Stein can have the same optimal shrinkage formula as a loss function that pertains to the covariance matrix such as Frobenius. Some loss functions, such as the three mentioned in the previous remark, are invariant to matrix inversion. In addition, it is possible to obtain an optimal shrinkage estimator of the covariance matrix with respect to Frobenius loss by shrinking the eigenvalues of the precision matrix. For all these reasons, one cannot ask whether we are shrinking the covariance matrix or the precision matrix because it is not a well-posed question. ■

5 Singular Case

We address the case $p > n$ by considering the inverses of the non-null sample eigenvalues only. The shrunken inverse eigenvalues are then given by:

$$\widehat{\delta}_{n,i}^{-1} := \begin{cases} \left(\dfrac{p}{n} - 1\right) \times \dfrac{1}{n} \displaystyle\sum_{j=p-n+1}^{p} \lambda_{n,j}^{-1} & i = 1, \ldots, p-n \\ \lambda_{n,i}^{-1} \, \mathcal{A}^2_{\underline{\widehat{\theta}}_n}\left(\lambda_{n,i}^{-1}\right) & i = p-n+1, \ldots, p \end{cases} \qquad \text{where} \tag{5.1}$$

$$\underline{\widehat{\theta}}_n(x) := \frac{1}{n} \sum_{j=p-n+1}^{p} \lambda_{n,j}^{-1} \frac{\lambda_{n,j}^{-1} - x}{\left(\lambda_{n,j}^{-1} - x\right)^2 + h_n^2 \lambda_{n,j}^{-2}} \qquad \text{and} \tag{5.2}$$

$$\mathcal{A}^2_{\underline{\widehat{\theta}}_n}(x) := \left[\frac{1}{n} \sum_{j=p-n+1}^{p} \lambda_{n,j}^{-1} \frac{\lambda_{n,j}^{-1} - x}{\left(\lambda_{n,j}^{-1} - x\right)^2 + h_n^2 \lambda_{n,j}^{-2}}\right]^2 + \left[\frac{1}{n} \sum_{j=p-n+1}^{p} \lambda_{n,j}^{-1} \frac{h_n \lambda_{n,j}^{-1}}{\left(\lambda_{n,j}^{-1} - x\right)^2 + h_n^2 \lambda_{n,j}^{-2}}\right]^2 . \tag{5.3}$$

Appendix E goes through the derivations in more detail. This shrinkage formula also works for Minimum Variance and Inverse Stein loss. It should be post-processed through (4.16) as before.

Note that Equations (5.2)–(5.3) are the same as their counterparts in Theorem 4.1, with the proviso that averaging only extends over non-null eigenvalues. Also, if we compare Equation (4.13) with the bottom line of Equation (5.1), the first two terms on the right-hand side inherited from Stein's loss have disappeared and all the weight has shifted onto the third term. We consider

this covariance matrix estimator to be the same as the one defined by Theorem 4.1, given that these two definitions cover two different domains: $p > n$ and $p < n$ respectively, so we can use the same symbol $\widehat{\Sigma}_n^{\text{QIS}}$ both times.

This approach cannot be used to estimate the covariance matrix under Stein's loss or the Symmetrized Kullback-Leibler divergence when $p > n$. This is because, as shown by Ledoit and Péché (2011, Theorem 5), adjusting null eigenvalues in these two cases requires estimating the population eigenvalues $(\tau_{n,1}, \ldots, \tau_{n,p})$, which is feasible only through numerical inversion of the QuEST function; see Ledoit and Wolf (2015, 2017). This feature is yet another advantage of quadratic-inverse shrinkage.

There is a deep mathematical reason for this. If $p > n$, the number of null sample eigenvalues remains constant as long as the number of null population eigenvalues is less than $p-n$. Thus, the first $p-n$ null population eigenvalues are essentially undetectable. Only if there are $p-n+1$ null population eigenvalues can we detect that there are indeed $p-n+1$ null population eigenvalues, since we observe $p-n+1$ null sample eigenvalues, and that is the only possible reason why. Thus, there is an initial zone where a few null population eigenvalues fly under the radar screen, followed by an 'avalanche effect' of revelation when all of a sudden there are too many of them. For this reason, any loss function that involves the inverse population covariance matrix should be avoided *at all cost* when $p > n$, and this includes both Stein's loss and the Symmetrized Kullback-Leibler divergence, as well as all the other loss functions that are invariant with respect to matrix inversion on the manifold of symmetric positive-definite matrices (tensors).

6 Smoothing Parameter and Concentration Ratio

The only guidance from theory about selecting the smoothing parameter h_n is that it must be of the form $Kn^{-\alpha}$ for some constants $K > 0$ and $\alpha \in (0, 2/5)$. However, this leaves open the question of dependency on $c_n := p/n$, the concentration ratio. Assumption 3.1 has it converging to some limit c bounded away from zero and infinity, so in theory the concentration is a constant that could appear anywhere in the formula. For example, we could have $h_n = K(c_n)\, n^{-\alpha}$, or even $h_n = K(c_n) \times [c_n n]^{-\alpha} = K(c_n)\, p^{-\alpha}$. In the absence of theoretical insight into this important question, we undertake a comprehensive numerical analysis to map out the terrain.

6.1 Test Bench

We cover six different scenarios for the population spectrum (indexed $i = 1, \ldots, 6$ below):

1. Σ_n is the identity, and the variates are distributed as a "Student" t distribution with 5 degrees of freedom.
2. High condition number: the population eigenvalues are uniformly distributed on the interval $[1, 30]$; the variates are Gaussian.
3. Spectral separation: half of the population eigenvalues are uniformly distributed on the interval $[1, 2]$, and the other half on $[10, 11]$; the variates are still Gaussian.

4. Upper spike: the first $p-1$ population eigenvalues are distributed as per the semi-circular law (Wigner, 1955) with support $[1, 5]$, and the top one is a "spike" that lies well above the bulk at the value 10; the variates are still Gaussian.
5. Lower spike: the top $p-1$ population eigenvalues are distributed as per the semi-circular law with support $[1, 5]$, and the smallest one is a "reverse spike" that lies well below the bulk at the value 0.25; the variates are still Gaussian.
6. Skewed: many small eigenvalues, few large ones; based on the Marčenko and Pastur (1967) law with parameter 1/2, which implies a condition number ≈ 33; the variates are still Gaussian.

These six scenarios are furthermore crossed with the two loss functions, Inverse Stein (Definition 4.1) and Minimum Variance (Definition 4.3), for a total of $6 \times 2 = 12$ combinations.

6.2 Monte Carlo Simulation Results

To vary the ratio p/n, we consider a collection of concentration ratios c defined as the tangents of angles in the set $\{10, 20, \ldots, 80\}$ (in degrees). To explore what happens for large and small sample sizes, we select 12 values for $\sqrt{pn}$ logarithmically spaced between 75 and 500. These two choices imply that p and n vary between a low of $75 \times \sqrt{\tan(10°)} \approx 31$ and a high of $500 \times \sqrt{\tan(80°)} \approx 1191$, a broad enough range. For each pair (p, n), we run 1,000 simulations.

For every (p, n) combination, for every population spectrum $i \in \{1, \ldots, 6\}$, and for every loss function $j \in \{\mathrm{MV}, \mathrm{IS}\}$, we determine numerically the optimal bandwidth $h_n^*(p, n, i, j)$. This is the value that minimizes the average of loss j across 1,000 Monte Carlo simulations run for the specification (p, n, i, j). Figure 4 displays the optimal bandwidth as a function of the matrix dimension and the sample size. The three axes are in logarithmic scale. Every point on the surface is the average of 12,000 numbers: 2 loss functions $\times$ 6 population spectra $\times$ 1,000 simulations.

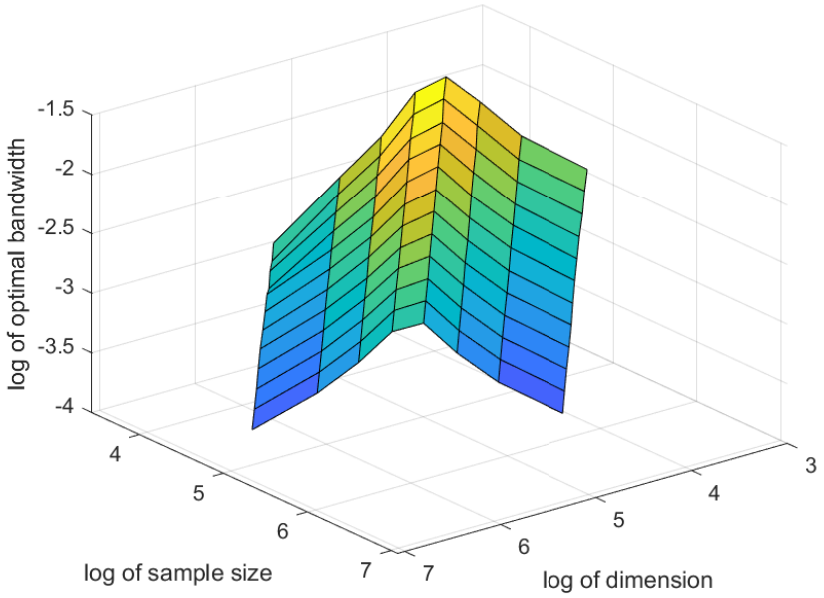

Figure 4: Dependence of the optimal bandwidth on the sample size and the matrix dimension.

The optimal bandwidth decreases in p and/or n as expected, but its level, governed by the multiplier K, has a very clear dependence on the concentration ratio: it is high when p is close to n, and low otherwise. This makes sense because concentrations close to one generate many eigenvalues close to zero, which are hard to handle. This inverted V-shape pattern is universal: It looks exactly the same even if we partial out the results by loss function, or by shape of population spectrum. On the latter point, Appendix F.4 has further justification.

6.3 Numerical Calibration of the Optimal Bandwidth

In order to formally model this structure, we run two regressions:

$$\log\left[h_n^*(p,n,i,j)\right] = a + b_1 \log\left[\min\left(\frac{p}{n},\frac{n}{p}\right)\right] + b_2 \log[n] + \varepsilon_{p,n,i,j} \tag{6.1}$$

$$\log\left[h_n^*(p,n,i,j)\right] = a + b_1 \log\left[\min\left(\frac{p}{n},\frac{n}{p}\right)\right] + b_2 \log[p] + \varepsilon_{p,n,i,j}\ . \tag{6.2}$$

Table 1 presents the results obtained with Matlab's `fitlm` routine.

Model	(6.1)	(6.2)
Intercept	−1.021	0.256
	(0.071)	(0.040)
$\log\big[\min(p/n, n/p)\big]$	0.659	0.659
	(0.018)	(0.010)
$\log[n]$	−0.149	
	(0.013)	
$\log[p]$		−0.392
		(0.007)
n	1152	1152
R^2	0.572	0.864

Table 1: Fitting linear models (6.1) and (6.2) to the optimal bandwidth in log-space.

One can clearly see that matrix dimension provides a better fit than sample size. Therefore, we are going into the direction of a bandwidth formula of the type $h_n = K(c_n)\, p^{-\alpha}$, rather than $h_n = K(c_n)\, n^{-\alpha}$. Looking at Panel B specifically, the R^2 is extremely high at 86.4%, which justifies our modeling choice for the dependency of K on c_n, with a symmetric drop away from $p = n$ in the directions $p < n$ and $p > n$. One important aspect is that the exponent of p is close to the -0.40 boundary from theory, so we suggest rounding it to -0.35. Further, the exponent of $\max(p/n, n/p)$ is quite close to twice this number: 0.70. Finally, the intercept is quite close to zero. Therefore, in the interest of elegance, our final recommendation for the bandwidth formula is simply:

$$h_n := \underbrace{\min\left(\frac{p^2}{n^2}, \frac{n^2}{p^2}\right)^{0.35}}_{K(c_n)} \times p^{-0.35}\,. \tag{6.3}$$

Based on more than a million Monte Carlo simulations, we consider this formula to be a safe all-around bandwidth that espouses the salient features of the problem at hand. Appendix F.4 gives supplementary evidence that our proposed estimator is not overly sensitive to the choice of smoothing hyper-parameter, as long as it remains in the general ballpark of Equation (6.3).

7 Numerical Performance of Quadratic-Inverse Shrinkage

Even though the main purpose of the present paper is essentially to exploit an unexpected connection between Stein's (1975) first-generation nonlinear shrinkage estimator of the covariance matrix and the latest advances in second-generation large-dimensional asymptotics, we still need to show how our proposed estimator — Quadratic-Inverse Shrinkage with smoothing parameter chosen as per Section 6 — performs relative to the current state-of-the-art. We put together a stable of seven competitors that do not assume any *a priori* information on the orientation of the eigenvectors of the true (unobservable) covariance matrix.

Sample The sample covariance matrix S_n.

Linear The linear shrinkage estimator of Ledoit and Wolf (2004).

NERCOME The nonlinear shrinkage estimator using cross-validation due to Lam (2016).

QIS The quadratic-inverse shrinkage estimator of Theorem 4.1 with smoothing parameter h_n chosen as per prescription (6.3).

Analytical The analytical nonlinear shrinkage formula of Ledoit and Wolf (2020, Section 4.7).

QuEST The nonlinear shrinkage estimator of Ledoit and Wolf (2015), which is based on numerical inversion of the QuEST function.

FSOPT The finite-sample optimal estimator $\overline{S}_n$ defined underneath Proposition 4.2, which would require knowledge of the unobservable population covariance matrix Σ_n, and thus is not applicable in the real world.

The first and the last are used for benchmarking purposes, as they generate the percentage relative improvement in average loss (PRIAL), defined for any estimator $\widehat{\Sigma}_n$ as

$$\text{PRIAL}_n\big(\widehat{\Sigma}_n\big) := \frac{\mathbb{E}\big[\mathcal{L}_n^{\text{FR}}\big(S_n, \Sigma_n\big)\big] - \mathbb{E}\big[\mathcal{L}_n^{\text{FR}}\big(\widehat{\Sigma}_n, \Sigma_n\big)\big]}{\mathbb{E}\big[\mathcal{L}_n^{\text{FR}}\big(S_n, \Sigma_n\big)\big] - \mathbb{E}\big[\mathcal{L}_n^{\text{FR}}\big(\overline{S}_n, \Sigma_n\big)\big]} \times 100\% \; . \tag{7.1}$$

The expectation $\mathbb{E}[\cdot]$ is in practice taken as the average across $\max\{100, \min\{1000, 10^5/p\}\}$ Monte Carlo simulations; for example, in dimension $p = 500$, we only need to run 200 simulations instead of 1000 to get reliable results.

From extant literature, we can already list some basic facts about these competitors.

- Linear shrinkage always beats the sample covariance matrix but, depending on parameter configurations, can either get very close to the FSOPT, or leave money on the table.
- The three nonlinear shrinkage estimators (NERCOME, Analytical, and QuEST) remain close to the FSOPT in any parameter configuration, so it is very hard to beat them in terms of accuracy.
- Sample, Linear, Analytical, and FSOPT have closed-form expressions, so it is very hard to beat them in terms of speed and scalability in ultra-high dimensions; however, the numerical estimators NERCOME and QuEST are orders of magnitude slower.

Therefore, we will be able to qualify the QIS estimator as 'state of the art' if it has similar accuracy to NERCOME, Analytical, and QuEST; and similar speed/scalability to Sample, Linear, Analytical, and FSOPT. Also worth pointing out is that two of these estimators, namely Analytical and QuEST, are 'outliers' in the present context because their formulas derive not from statistics but from the Random Matrix Theory invented by 1963 Physics Nobel prize winner Eugene Wigner (1955) to model the wave functions of quantum mechanical systems, by way of the subsequent work of Marčenko and Pastur (1967) and their successors.

Even though the Monte Carlo simulations in this section focus on the Frobenius loss, because it has been widely accepted across many applied fields over the past couple of decades, supplementary simulations in Appendix F.3 indicate that similar conclusions would carry over to the Inverse Stein's loss and the Minimum Variance loss.

7.1 Baseline Scenario

The simulations are organized around a baseline scenario. Each parameter will be subsequently varied to assess the robustness of the conclusions. The baseline scenario is:

- the matrix dimension is $p = 200$;
- the sample size is $n = 600$; therefore, the concentration ratio p/n is equal to $1/3$;
- the condition number of the population covariance matrix is 10;
- 20% of population eigenvalues are equal to 1, 40% are equal to 3, and 40% are equal to 10;
- and the variates are normally distributed.

The distribution of the population eigenvalues is a particularly interesting and difficult case introduced and analyzed in detail by Bai and Silverstein (1998). We have purposefully selected a shape of population spectrum left untouched by the calibration round in Section 6.

Table 2 presents estimator performances under the baseline scenario. Computational times come from a 3.3GHz Mac Pro desktop computer running Matlab R2020a.

Estimator	Sample	Linear	NERCOME	QIS	Analytical	QuEST	FSOPT
Average Loss	39.1	23.5	17.1	16.8	16.6	16.2	16.1
PRIAL	0%	68%	96%	97%	98%	99%	100%
Time (ms)	< 1	1	2,117	3	3	1,644	3

Table 2: Simulation results for the baseline scenario.

The 0% PRIAL for the sample covariance matrix and the 100% PRIAL for the finite-sample optimal estimator are by construction. Linear shrinkage performs well but leaves some money on the table. The four nonlinear shrinkage formulas deliver near-perfect performance in the 96%+ range, with NERCOME and QuEST being much slower by orders of magnitude, as expected.

7.2 Convergence

Under large-dimensional asymptotics, the matrix dimension p and the sample size n go to infinity together, while their ratio p/n converges to some limit c. In the first experiment, p and n increase together, with their ratio fixed at the baseline value of $1/3$. Figure 5 displays the results.

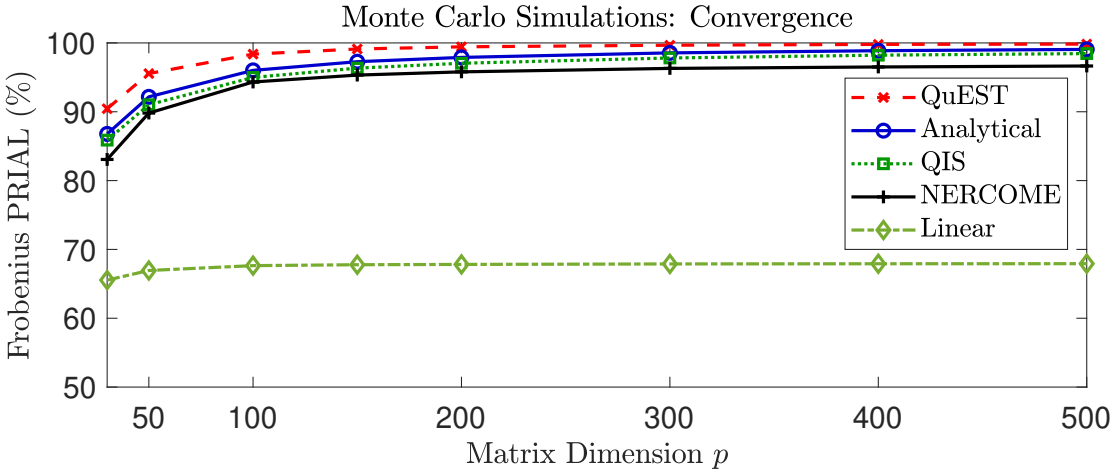

Figure 5: Evolution of the PRIAL as the matrix dimension and the sample size go to infinity together.

The four nonlinear shrinkage methods perform approximately the same as one another. They do well even in small dimensions, but do better as the dimension grows large. The difference between the PRIALs of Analytical and QIS is never more than 1%, which is very small.

Remark 7.1. Analytical, QIS and QuEST have shrinkage functions that behave essentially the same under large-dimensional asymptotics. They have the same limiting loss. Therefore, we should expect their performances to be nearly identical for large (p, n). For small and moderate (p, n), we would expect the performance of QuEST to be somewhat better compared to both Analytical and QIS, since it exploits the feature that f is the density of a limiting sample spectral distribution that is the output of the Fundamental Equation of Random Matrix Theory; hence, QuEST can be considered a model-based estimator. By contrast, Analytical and QIS do not exploit this feature of f, and thus can be considered model-free estimators. ■

To see what happens when matrices become very large, we consider the case $p = 10,000$ in Table 3. At this level, the numerical methods QuEST and NERCOME can no longer follow, even with a powerful computer, so we only consider the other estimators.

Estimator	Sample	Linear	QIS	Analytical	FSOPT
Average Loss	38.89	23.37	16.12	16.08	16.07
PRIAL	0%	68.0%	99.8%	99.9%	100%
Time (s)	5	10	31	32	35

Table 3: Results of 100 Monte Carlo simulations for $p = 10,000$ and $n = 30,000$.

One can see that letting dimension go to infinity does nothing for linear shrinkage, but it brings nonlinear shrinkage ever-closer to the maximum allowable level of improvement, 100%. QIS is

slightly worse than Analytical, but the difference is too small to be material. In unreported simulations, we even managed to go up to $p = 20,000$ and $n = 60,000$. The average computational time increased to just under four minutes per estimator, and was mostly devoted to extracting the sample covariance eigenvalues and eigenvectors.

7.3 Concentration Ratio

We vary the concentration ratio p/n from 0.1 to 0.9 while holding the product $p \times n$ constant at the level it had under the baseline scenario, namely, $p \times n = 120,000$. (In this way, we keep the amount of total information fixed, as measured by the number of entries in the matrix Y_n of Assumption 3.2b, as the concentration ratio varies.) Figure 6 displays the resulting PRIALs.

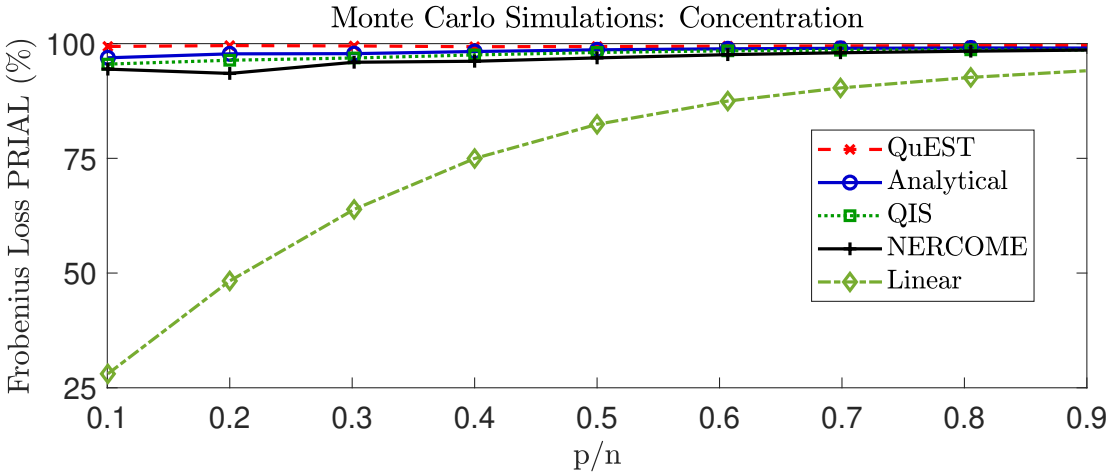

Figure 6: Evolution of the PRIAL of various estimators as a function of the ratio of the matrix dimension to the sample size.

Linear shrinkage performs very well in high concentrations but less so in low concentrations, even though it still improves decisively over the sample covariance matrix across the board, as evidenced by its strictly positive PRIALs. The four nonlinear shrinkage methods perform approximately the same as one another, with QuEST remaining the gold standard, and the others performing nearly as well (for any practical purposes).

7.4 Condition Number

We start again from the baseline scenario and, this time, vary the condition number of the population covariance matrix, called κ. We set 20% of the population eigenvalues equal to 1, 40% equal to $(2\kappa + 7)/9$, and 40% equal to κ. Thus, the baseline scenario corresponds to $\kappa = 10$. In this experiment, we let κ vary from $\kappa = 3$ to $\kappa = 30$. This corresponds to linearly squeezing or stretching the distribution of population eigenvalues. Figure 7 displays the resulting PRIALs.

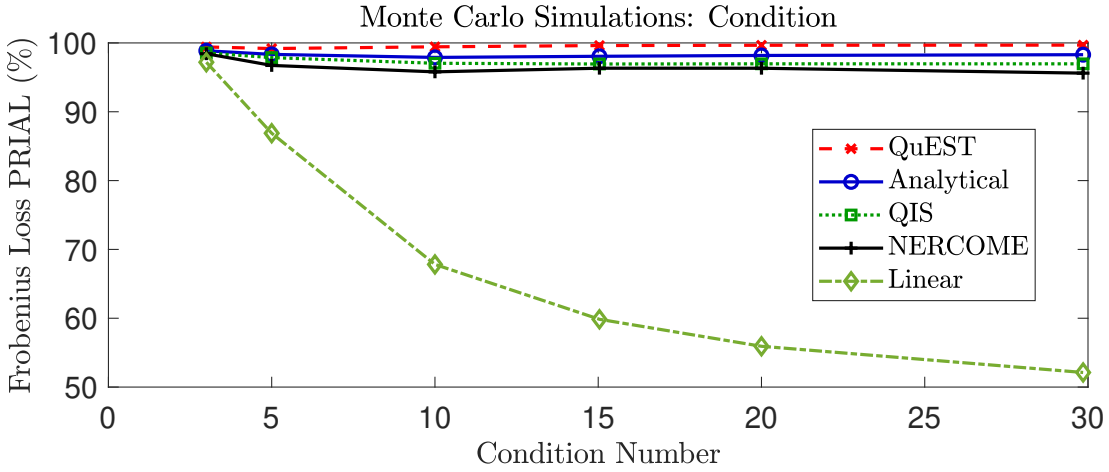

Figure 7: Evolution of the PRIAL of various estimators as a function of the condition number of the population covariance matrix.

Linear shrinkage performs very well for low condition numbers, but leaves some money on the table when eigenvalues are dispersed, as predicted theoretically by Ledoit and Wolf (2004, Figure 5). The nonlinear shrinkage formulas capture nearly all the potential for loss reduction.

7.5 Non-Normality

In this experiment, we start from the baseline scenario and change the distribution of the variates. We study the Bernoulli coin toss distribution, which is the most platykurtic of all distributions, the Laplace distribution, which is leptokurtotic, and the Student t-distribution with 5 degrees of freedom, also leptokurtotic. All of these are suitably normalized to have mean zero and variance one, if necessary. Table 4 presents the results.

Distribution	Linear	NERCOME	QIS	Analytical	QuEST
Bernoulli	67.5%	95.9%	97.5%	98.1%	99.4%
Laplace	68.4%	95.7%	96.4%	97.5%	99.1%
Student t_5	68.7%	95.7%	95.4%	96.5%	98.2%

Table 4: Simulation results for various variate distributions (PRIAL).

This experiment confirms that results are not sensitive to the distribution of the variates.

7.6 Shape of the Distribution of Population Eigenvalues

Relative to the baseline scenario, we now move away from the clustered distribution for the population eigenvalues and try continuous distributions from the Beta family. They are linearly shifted and stretched so that the support is $[1, 10]$. A graphical illustration of the densities of the various Beta shapes is in Ledoit and Wolf (2018, Figure 8.4). Table 5 presents the results.

Beta Parameters	Linear	NERCOME	QIS	Analytical	QuEST
$(1, 1)$	92.8%	98.3%	98.5%	98.8%	99.2%
$(1, 2)$	96.6%	97.6%	98.4%	98.6%	98.8%
$(2, 1)$	97.2%	99.2%	98.9%	99.2%	99.6%
$(1.5, 1.5)$	96.1%	98.6%	98.7%	99.0%	99.3%
$(0.5, 0.5)$	82.8%	97.8%	98.2%	98.5%	99.1%
$(5, 5)$	99.1%	99.5%	99.0%	99.3%	99.7%
$(5, 2)$	99.0%	99.6%	99.0%	99.4%	99.8%
$(2, 5)$	98.4%	98.6%	98.7%	99.0%	99.2%

Table 5: Simulation results for various distributions of the population eigenvalues (PRIAL).

This time, linear shrinkage does much better overall, except perhaps for the bimodal shape $(0.5, 0.5)$. This is due to the fact that, in the other cases, the optimal nonlinear shrinkage formula happens to be almost linear. Nonlinear shrinkage formulas capture a nearly perfect percentage of the potential for variance reduction in all cases.

7.7 Singular Case

Finally, we run a counterpart of the Monte Carlo simulations in Section 7.3 for the case $c > 1$. We vary the concentration ratio p/n from 1.1 to 10 while holding the product $p \times n$ constant at the level it had under the baseline scenario, namely, $p \times n = 120,000$. (In this way, we keep the amount of total information fixed, as measured by the number of entries in the matrix Y_n of Assumption 3.2b, as the concentration ratio varies.) Figure 8 displays the resulting PRIALs.

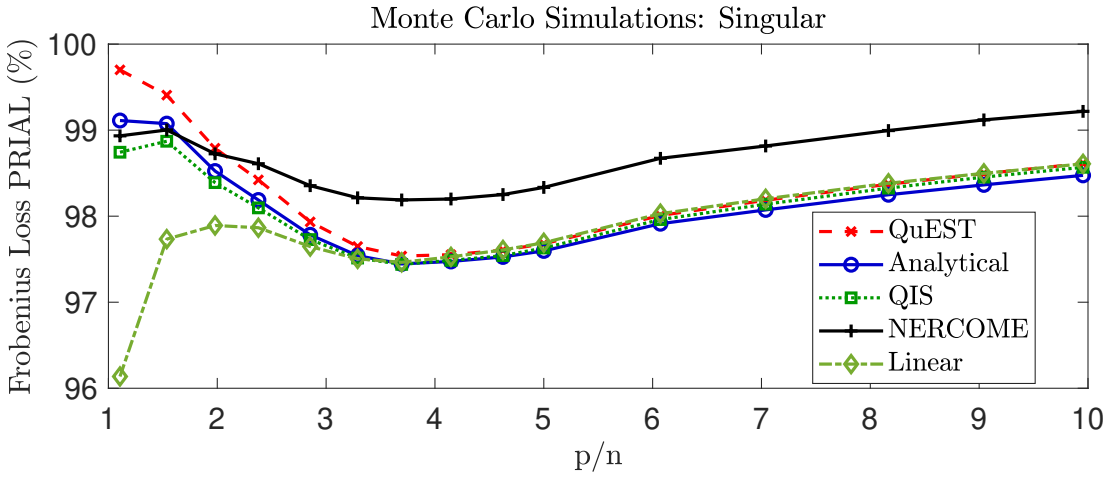

Figure 8: Frobenius PRIAL when the matrix dimension exceeds the sample size.

We draw the attention of the reader to the vertical scale of the figure: It starts at 96%. This confirms the trend that could be inferred from Figure 6: Higher concentration ratios make

all shrinkage estimators look good. At this level of performance, the exact ordering becomes relatively less important; overall, NERCOME does best.

7.8 Supplementary Monte Carlo Simulations

Because of space contraints, some further batches of numerical simulations are relegated to Appendix F. Their conclusions are briefly summarized below for convenience:

- There is no real need to post-process the QIS estimator with a numerical algorithm that would restore the order of the shrunken eigenvalues.
- Choosing a smoothing kernel different from the Cauchy density would only increase complexity without increasing performance.
- The general pattern of simulation results presented in Section 7 for the Frobenius loss is similar for the Inverse Stein's loss and the Minimum Variance loss.
- The performance of the QIS estimator is not sensitive to the specification of the smoothing parameter h_n, as long as we remain in the general area of our proposal from Section 6.
- The classic estimator of Stein (1975, 1977, 1986) is still hard to beat on its own terms, that is, with respect to Stein's loss. The value of the QIS estimator is that it extends the same logic to alternative loss functions that are harder to handle mathematically, and potentially more attractive in practice (such as Frobenius loss and Minimum Variance loss), as well as being able to handle the challenging $p > n$ case.

7.9 Overall Comparison of Performance Results

In terms of accuracy, the QIS estimator matches the high-water mark set by NERCOME, Analytical, and QuEST in hugging close to the FSOPT no matter what happens — unlike linear shrinkage, whose percentage improvement (albeit always positive) can fluctuate according to parameter configurations. In terms of speed and of scalability into ultra-high dimensions, the QIS estimator is in the efficient group alongside Sample, Linear, Analytical and FSOPT — a decisive advantage over the numerical methods NERCOME and QuEST.

If we define 'state-of-the-art' as close-to-FSOPT accuracy across the parameter space, and a scalable closed-formed mathematical expression, then QIS the only such estimator apart from Analytical, although there could conceivably be more invented by future researchers over the course of scientific progress. What makes QIS special in this class so far is that its formula

- originates from statistical decision theory (Stein, 1975), whereas the Analytical shrinkage formula originates in the physics of random matrix theory (Marčenko and Pastur, 1967);
- is intelligible because it is a simple adaptation of the Stein shrinker, which visibly attracts sample eigenvalues to close neighbors on either side, decaying with distance;
- and has lower degree of complexity because it is second-order (quadratic) shrinkage, as opposed to infinite-order nonlinear.

8 Conclusion

Stein's (1975,1977,1986) seminal work has garnered a lot of attention over the years from researchers interested in estimating covariance matrices of dimension larger than three. It is hard to make an original contribution on top of such a body of knowledge, but we (i) reinterpret Stein's ostensibly nonlinear shrinkage formula as linear in inverse-eigenvalues space; (ii) smooth out his shrinker to make it continuous instead of divergent; and (iii) address more practically-oriented loss functions that work even when variables outnumber observations, by adjoining a quadratic component.

Given that this construct harnesses the latest techniques in large-dimensional asymptotic theory, we believe that it is not just tying up loose ends from the past, but also the foundation for a new covariance matrix estimator that will prove useful to future researchers. The relentless search for simpler formulas is a priority for the community because future developments will be easier to build on top of transparent insights instead of arcane ones. The intimate connection we established between one nonlinear shrinkage formula from the first generation (finite samples) and another nonlinear shrinkage formula from the second generation (large-dimensional asymptotics) is quite unexpected due to the wide generation gap in techniques and methodology, so the most likely reason is that a deeper mathematical truth has, at least partially, been unearthed.

References

Abrahamsson, R., Selen, Y., and Stoica, P. (2007). Enhanced covariance matrix estimators in adaptive beamforming. In *2007 IEEE International Conference on Acoustics, Speech and Signal Processing – ICASSP 2007*, volume II, pages 969–972.

Arsigny, V., Fillard, P., Pennec, X., and Ayache, N. (2006). Log-euclidean metrics for fast and simple calculus on diffusion tensors. *Magnetic Resonance in Medicine*, 56(2):411–421.

Ayer, M., Brunk, H. D., Ewing, G. M., Reid, W. T., Silverman, E., et al. (1955). An empirical distribution function for sampling with incomplete information. *Annals of Mathematical Statistics*, 26(4):641–647.

Bai, Z. D. and Silverstein, J. W. (1998). No eigenvalues outside the support of the limiting spectral distribution of large-dimensional random matrices. *Annals of Probability*, 26(1):316–345.

Bickel, P. J. and Levina, E. (2008a). Covariance regularization by thresholding. *Annals of Statistics*, 36(6):2577–2604.

Bickel, P. J. and Levina, E. (2008b). Regularized estimation of large covariance matrices. *Annals of Statistics*, 36(1):199–227.

Bodnar, T., Gupta, A. K., and Parolya, N. (2016). Direct shrinkage estimation of large dimensional precision matrix. *Journal of Multivariate Analysis*, 146:223–236. Special Issue on Statistical Models and Methods for High or Infinite Dimensional Spaces.

Capon, J. (1969). High-resolution frequency-wavenumber spectrum analysis. *Proceedings of the IEEE*, 57(8):1408–1418.

Chen, S. X., Zhang, L.-X., and Zhong, P.-S. (2010). Tests for high-dimensional covariance matrices. *Journal of the American Statistical Association*, 105(490):810–819.

Chen, Y., Wiesel, A., and Hero, A. O. (2009). Shrinkage estimation of high dimensional covariance matrices. IEEE International Conference on Acoustics, Speech, and Signal Processing, Taiwan.

Chen, Y., Wiesel, A., and Hero, A. O. (2011). Robust shrinkage estimation of high-dimensional covariance matrices. *IEEE Transactions on Signal Processing*, 59(9):4097–4107.

Daniels, M. J. and Kass, R. E. (2001). Shrinkage estimators for covariance matrices. *Biometrics*, 57:1173–1184.

Deligianni, F., Centeno, M., Carmichael, D. W., and Clayden, J. D. (2014). Relating resting-state fMRI and EEG whole-brain connectomes across frequency bands. *Frontiers in Neuroscience*, 8:258.

Dey, D. K. and Srinivasan, C. (1985). Estimation of a covariance matrix under Stein's loss. *Annals of Statistics*, 13(4):1581–1591.

Donoho, D. L., Gavish, M., and Johnstone, I. M. (2018). Optimal shrinkage of eigenvalues in the spiked covariance model. *Annals of Statistics*, 46(4):1742–1778.

Efron, B. and Morris, C. (1976). Multivariate empirical bayes and estimation of covariance matrices. *Annals of Statistics*, 4(1):22–32.

El Karoui, N. (2008). Operator norm consistent estimation of large-dimensional sparse covariance matrices. *Annals of Statistics*, 36(6):2717–2756.

Elsheikh, A. H., Wheeler, M. F., and Hoteit, I. (2013). An iterative stochastic ensemble method for parameter estimation of subsurface flow models. *Journal of Computational Physics*, 242:696–714.

Engle, R. F. and Colacito, R. (2006). Testing and valuing dynamic correlations for asset allocation. *Journal of Business & Economic Statistics*, 24(2):238–253.

Engle, R. F., Ledoit, O., and Wolf, M. (2019). Large dynamic covariance matrices. *Journal of Business & Economic Statistics*, 37(2):363–375.

Epanechnikov, V. A. (1969). Non-parametric estimation of a multivariate probability density. *Theory of Probability & Its Applications*, 14(1):153–158.

Förstner, W. and Moonen, B. (1999). A metric for covariance matrices. In Krumm, F. and Schwarze, V. S., editors, *Quo vadis geodesia ...? Festschrift for Erik W. Grafarend on the occasion of his 60th birthday*, number 1999.6 in Technical Reports of the Department of Geodesy and Geoinformatics, pages 113–128. University of Stuttgart, Institute of Geodesy.

Gabor, D. (1946). Theory of communication. part 1: The analysis of information. *Journal of the Institution of Electrical Engineers*, 93(26):429–441.

Guo, S.-M., He, J., Monnier, N., Sun, G., Wohland, T., and Bathe, M. (2012). Bayesian approach to the analysis of fluorescence correlation spectroscopy data II: Application to simulated and in vitro data. *Analytical Chemistry*, 84(9):3880–3888.

Haff, L. (1979). Estimation of the inverse covariance matrix: Random mixtures of the inverse wishart matrix and the identity. *Annals of Statistics*, pages 1264–1276.

Haff, L. R. (1980). Empirical Bayes estimation of the multivariate normal covariance matrix. *Annals of Statistics*, 8:586–597.

Haff, L. R. (1991). The variational form of certain Bayes estimators. *Annals of Statistics*, 19(3):1163–1190.

Hasselmann, K. (1993). Optimal fingerprints for the detection of time-dependent climate change. *Journal of Climate*, 6(10):1957–1971.

IPCC (2007). Climate change 2007: the physical science basis. In Solomon, S., Qin, D., Manning, M., Marquis, M., Averyt, K., Tignor, M. M., Miller, H. L., and Chen, Z., editors, *Working Group I Contribution to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*, volume 4. Cambridge University Press, Cambridge and New York. p. 996.

James, W. and Stein, C. (1961). Estimation with quadratic loss. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability* **1**, pages 361–380.

Jing, B.-Y., Pan, G., Shao, Q.-M., and Zhou, W. (2010). Nonparametric estimate of spectral density functions of sample covariance matrices: A first step. *Annals of Statistics*, 38(6):3724–3750.

John, S. (1971). Some optimal multivariate tests. *Biometrika*, 58(1):123–127.

Korniotis, G. M. (2008). Habit formation, incomplete markets, and the significance of regional risk for expected returns. *The Review of Financial Studies*, 21(5):2139–2172.

Krantz, S. G. (2009). *Explorations in Harmonic Analysis*. Birkhäuser, Boston.

Krishnamoorthy, K. and Gupta, A. (1989). Improved minimax estimation of a normal precision matrix. *Canadian Journal of Statistics*, 17(1):91–102.

Lam, C. (2016). Nonparametric eigenvalue-regularized precision or covariance matrix estimator. *Annals of Statistics*, 44(3):928–953.

Ledoit, O. and Péché, S. (2011). Eigenvectors of some large sample covariance matrix ensembles. *Probability Theory and Related Fields*, 150(1–2):233–264.

Ledoit, O. and Wolf, M. (2002). Some hpothesis tests for the covariance matrix when the dimension is large compared to the sample size. *Annals of Statistics*, 30(4):1081–1102.

Ledoit, O. and Wolf, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88(2):365–411.

Ledoit, O. and Wolf, M. (2012). Nonlinear shrinkage estimation of large-dimensional covariance matrices. *Annals of Statistics*, 40(2):1024–1060.

Ledoit, O. and Wolf, M. (2015). Spectrum estimation: A unified framework for covariance matrix estimation and PCA in large dimensions. *Journal of Multivariate Analysis*, 139(2):360–384.

Ledoit, O. and Wolf, M. (2017). Numerical implementation of the QuEST function. *Computational Statistics & Data Analysis*, 115:199–223.

Ledoit, O. and Wolf, M. (2018). Optimal estimation of a large-dimensional covariance matrix under Stein's loss. *Bernoulli*, 24(4B). 3791–3832.

Ledoit, O. and Wolf, M. (2020). Analytical nonlinear shrinkage of large-dimensional covariance matrices. *Annals of Statistics*, 40(5):3043–3065.

Lin, S. P. and Perlman, M. D. (1985). A Monte Carlo comparison of four estimators of a covariance matrix. In Krishnaiah, P. R., editor, *Multivariate Analysis VI: Proceedings of the International Symposium on Multivariate Analysis held at the University of Pittsburgh*, pages 411–429. North Holland, Amsterdam.

Loh, W.-L. (1991). Estimating covariance matrices. *Annals of Statistics*, 19(1):283–296.

Marčenko, V. A. and Pastur, L. A. (1967). Distribution of eigenvalues for some sets of random matrices. *Sbornik: Mathematics*, 1(4):457–483.

Markon, K. (2010). Modeling psychopathology structure: A symptom-level analysis of axis I and II disorders. *Psychological Medicine*, 40(2):273–288.

Markowitz, H. (1952). Portfolio selection. *Journal of Finance*, 7:77–91.

Moakher, M. and Batchelor, P. G. (2006). Symmetric positive-definite matrices: From geometry to applications and visualization. In Weickert, J. and Hagen, H., editors, *Visualization and Processing of Tensor Fields*, pages 285–298. Springer-Verlag, Berlin.

Nagao, H. (1973). On some test criteria for covariance matrix. *Annals of Statistics*, 4(1):700–709.

Pal, N. (1993). Estimating the normal dispersion matrix and the precision matrix from a decision-theoretic point of view: a review. *Statistical Papers*, 34(1):1–26.

Poularikas, A. D., editor (1998). *The Handbook of Formulas and Tables for Signal Processing*. CRC Press, Boca Raton.

Pyeon, D., Newton, M., Lambert, P., Den Boon, J., Sengupta, S., Marsit, C., Woodworth, C., Connor, J., Haugen, T., Smith, E., Kelsey, K., Turek, L., and Ahlquist, P. (2007). Fundamental differences in cell cycle deregulation in human papillomavirus-positive and human papillomavirus-negative head/neck and cervical cancers. *Cancer Research*, 67(10):4605–4619.

Rajaratnam, B. and Vincenzi, D. (2016a). A note on covariance estimation in the unbiased estimator of risk framework. *Journal of Statistical Planning and Inference*, 175:25–39.

Rajaratnam, B. and Vincenzi, D. (2016b). A theoretical study of Stein's covariance estimator. *Biometrika*, 103(3):653–666.

Ribes, A., Azaïs, J.-M., and Planton, S. (2009). Adaptation of the optimal fingerprint method for climate change detection using a well-conditioned covariance matrix estimate. *Climate Dynamics*, 33(5):707–722.

Schäfer, J. and Strimmer, K. (2005). A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical Applications in Genetics and Molecular Biology*, 4(1). Article 32.

Silverstein, J. W. (1995). Strong convergence of the empirical distribution of eigenvalues of large-dimensional random matrices. *Journal of Multivariate Analysis*, 55:331–339.

Silverstein, J. W. and Bai, Z. D. (1995). On the empirical distribution of eigenvalues of a class of large-dimensional random matrices. *Journal of Multivariate Analysis*, 54:175–192.

Silverstein, J. W. and Choi, S. I. (1995). Analysis of the limiting spectral distribution of large-dimensional random matrices. *Journal of Multivariate Analysis*, 54:295–309.

Stein, C. (1975). Estimation of a covariance matrix. Rietz lecture, 39th Annual Meeting IMS. Atlanta, Georgia.

Stein, C. (1977). Lectures on the theory of estimation of many parameters (in Russian). In Ibragimov, I. A. and Nikulin, M. S., editors, *Studies in the Statistical Theory of Estimation, Part I*, volume 74 of *Proceedings of Scientific Seminars of the Steklov Institute, Leningrad Division*, pages 4–65.

Stein, C. (1986). Lectures on the theory of estimation of many parameters. *Journal of Mathematical Sciences*, 34(1):1373–1403.

Stoica, P., Li, J., Zhu, X., and Guerci, J. R. (2008). On using a priori knowledge in space-time adaptive processing. *IEEE Transactions on Signal Processing*, 56(6):2598–2602.

Tenenhaus, A. and Tenenhaus, M. (2011). Regularized generalized canonical correlation analysis. *Psychometrika*, 76(2):257.

Titchmarsh, E. C. (1948). *Introduction to the Theory of Fourier Integrals*. Clarendon Press, Oxford, second edition.

Tsukuma, H. (2005). Estimating the inverse matrix of scale parameters in an elliptically contoured distribution. *Journal of the Japan Statistical Society*, 35(1):21–39.

Vakman, D. (1996). On the analytic signal, the Teager-Kaiser energy algorithm, and other methods for defining amplitude and frequency. *IEEE Transactions on Signal Processing*, 44(4):791–797.

Vidaurre, C., Krämer, N., Blankertz, B., and Schlögl, A. (2009). Time domain parameters as a feature for EEG-based brain–computer interfaces. *Neural Networks*, 22(9):1313–1319.

Wigner, E. P. (1955). Characteristic vectors of bordered matrices with infinite dimensions. *Annals of Mathematics*, 62(3):548–564.

Won, J.-H., Lim, J., Kim, S.-J., and Rajaratnam, B. (2013). Condition-number regularized covariance estimation. *Journal of the Royal Statistical Society, Series B*, 75(3):427–450.

Yang, R. and Berger, J. O. (1994). Estimation of a covariance matrix using the reference prior. *Annals of Statistics*, 22(3):1195–1211.

A Programming Code

The Matlab function for the Quadratic-Inverse Shrinkage estimator of the covariance matrix has only 20 or so lines of documented code, which makes for easy understanding and customization.

```
function sigmahat=QIS(Y,k)          % sigmahat:covariance matrix; Y:raw data
%%% EXTRACT sample eigenvalues sorted in ascending order and eigenvectors %%%
[N,p]=size(Y);                      % sample size and matrix dimension
if (nargin<2)||isnan(k)||isempty(k) % default setting
   Y=Y-repmat(mean(Y),[N 1]);       % demean the raw data matrix
   k=1;                             % subtract one degree of freedom
end
n=N-k;                              % adjust effective sample size
c=p/n;                              % concentration ratio
sample=(Y'*Y)./n;                   % sample covariance matrix
[u,lambda]=eig(sample,'vector');    % spectral decomposition
[lambda,isort]=sort(lambda);        % sort eigenvalues in ascending order
u=u(:,isort);                       % eigenvectors follow their eigenvalues
%%% COMPUTE Quadratic-Inverse Shrinkage estimator of the covariance matrix %%%
h=min(c^2,1/c^2)^0.35/p^0.35;              % smoothing parameter
invlambda=1./lambda(max(1,p-n+1):p);       % inverse of (non-null) eigenvalues
Lj=repmat(invlambda,[1 min(p,n)])';        % like  1/lambda_j
Lj_i=Lj-Lj';                               % like (1/lambda_j)-(1/lambda_i)
theta=mean(Lj.*Lj_i./(Lj_i.^2+h^2.*Lj.^2),2);     % smoothed Stein shrinker
Htheta=mean(Lj.*(h.*Lj)./(Lj_i.^2+h^2.*Lj.^2),2); % its conjugate
Atheta2=theta.^2+Htheta.^2;                        % its squared amplitude
if p<=n % case where sample covariance matrix is not singular
   delta=1./((1-c)^2*invlambda+2*c*(1-c)*invlambda.*theta ...
      +c^2*invlambda.*Atheta2);            % optimally shrunk eigenvalues
else    % case where sample covariance matrix is singular
   delta0=1./((c-1)*mean(invlambda));      % shrinkage of null eigenvalues
   delta=[repmat(delta0,[p-n 1]);1./(invlambda.*Atheta2)];
end
deltaQIS=delta.*(sum(lambda)/sum(delta)); % preserve trace
sigmahat=u*diag(deltaQIS)*u';             % reconstruct covariance matrix
```

The QIS function transforms an $n \times p$ matrix `Y` containing n i.i.d. samples of p variables into the $p \times p$ nonlinear shrinkage covariance matrix estimator `sigmahat`. If the second (optional) parameter k is absent, not-a-number, or empty, then the algorithm demeans the data by default, and adjusts the effective sample size accordingly. If the user inputs $k = 0$, then no demeaning takes place; if (s)he inputs $k = 1$, then it signifies that the data Y has already been demeaned.

We also have Python code that delivers the same results (verified for Python version 3.7.9):

```
#Imports
import numpy as np
import pandas as pd
import math

#Sigmahat function
def sigmahat(Y,k=None):
    #Pre-Conditions: Y is a valid pd.dataframe and optional arg- k which can be
    #    None, np.nan or int
    #Post-Condition: Sigmahat dataframe is returned

    #Set df dimensions
    N = Y.shape[0]                                          #num of columns
    p = Y.shape[1]                                             #num of rows

    #default setting
    if (k is None or math.isnan(k)):
        Y = Y.sub(Y.mean(axis=0), axis=1)                                  #demean
        k = 1

    #vars
    n = N-k                                          # adjust effective sample size
    c = p/n                                                  # concentration ratio

    #Cov df: sample covariance matrix
    sample = pd.DataFrame(np.matmul(Y.T.to_numpy(),Y.to_numpy()))/n
    sample = (sample+sample.T)/2                                  #make symmetrical

    #Spectral decomp
    lambda1, u = np.linalg.eigh(sample)             #use Cholesky factorisation
    #                                                 based on hermitian matrix
    lambda1 = lambda1.real.clip(min=0)               #reset negative values to 0
    dfu = pd.DataFrame(u,columns=lambda1)    #create df with column names lambda
    #                                                 and values u
    dfu.sort_index(axis=1,inplace = True)                #sort df by column index
    lambda1 = dfu.columns                                #recapture sorted lambda

    #COMPUTE Quadratic-Inverse Shrinkage estimator of the covariance matrix
    h = (min(c**2,1/c**2)**0.35)/p**0.35                      #smoothing parameter
```

```
    invlambda = 1/lambda1[max(1,p-n+1)-1:p]  #inverse of (non-null) eigenvalues
    dfl = pd.DataFrame()
    dfl['lambda'] = invlambda
    Lj = dfl[np.repeat(dfl.columns.values,min(p,n))]          #like  1/lambda_j
    Lj = pd.DataFrame(Lj.to_numpy())                         #Reset column names
    Lj_i = Lj.subtract(Lj.T)                      #like (1/lambda_j)-(1/lambda_i)

    theta = Lj.multiply(Lj_i).div(Lj_i.multiply(Lj_i).add(
        Lj.multiply(Lj)*h**2)).mean(axis = 0)          #smoothed Stein shrinker
    Htheta = Lj.multiply(Lj*h).div(Lj_i.multiply(Lj_i).add(
        Lj.multiply(Lj)*h**2)).mean(axis = 0)                    #its conjugate
    Atheta2 = theta**2+Htheta**2                           #its squared amplitude

    if p<=n:               #case where sample covariance matrix is not singular
         delta = 1 / ((1-c)**2*invlambda+2*c*(1-c)*invlambda*theta \
                      +c**2*invlambda*Atheta2)    #optimally shrunk eigenvalues
         delta = delta.to_numpy()
    else:
        delta0 = 1/((c-1)*np.mean(invlambda.to_numpy())) #shrinkage of null
        #                                                 eigenvalues
        delta = np.repeat(delta0,p-n)
        delta = np.concatenate((delta, 1/(invlambda*Atheta2)), axis=None)

    deltaQIS = delta*(sum(lambda1)/sum(delta))                  #preserve trace

    temp1 = dfu.to_numpy()
    temp2 = np.diag(deltaQIS)
    temp3 = dfu.T.to_numpy().conjugate()
    #reconstruct covariance matrix
    sigmahat = pd.DataFrame(np.matmul(np.matmul(temp1,temp2),temp3))
    return sigmahat
```

One key point is that the algorithm that decomposes the sample covariance matrix into eigenvectors and eigenvalues must be explicitly told that the sample covariance matrix is symmetric. Matlab detects this automatically, but within Python's `numpy` library, we must call the tailor-made `linalg.eigh` (based on the Cholesky decomposition) instead of the usual `linalg.eig` function.[6] Otherwise, in the $p > n$ case, the procedure tends to return complex eigenvalues and eigenvectors, which would be impossible to handle. We flag this out in the interest of future researchers who wish to implement the QIS estimator in other languages.

[6]This issue was unearthed and resolved by our Matlab-to-Python translator Patrick Ledoit.

B Proofs of Propositions

B.1 Proof of Proposition 2.1

Let $\widetilde{\Delta}_n =: \mathsf{Diag}\big(\widetilde{\delta}_{n,1}, \ldots, \widetilde{\delta}_{n,p}\big)$. Then

$$\begin{aligned}
\mathcal{L}_n^{\mathrm{ST}}\left(\Sigma_n, U_n\widetilde{\Delta}_n U_n'\right) &= \frac{1}{p}\mathsf{Tr}\Big(\Sigma_n^{-1}U_n\widetilde{\Delta}_n U_n'\Big) - \frac{1}{p}\log\,\det\Big(\Sigma_n^{-1}U_n\widetilde{\Delta}_n U_n'\Big) - 1 \\
&= \frac{1}{p}\mathsf{Tr}\Big(U_n'\Sigma_n^{-1}U_n\widetilde{\Delta}_n\Big) - \frac{1}{p}\log\,\det\Big(\Sigma_n^{-1}\widetilde{\Delta}_n\Big) - 1 \\
&= \frac{1}{p}\sum_{i=1}^{p}\Big[u_{n,i}'\Sigma_n^{-1}u_{n,i}\cdot\widetilde{\delta}_{n,i} - \log\big(\widetilde{\delta}_{n,i}\big)\Big] + \text{constant}\ .
\end{aligned}$$

The first-order condition is $u_{n,i}'\Sigma_n^{-1}u_{n,i}^{-1} - \widetilde{\delta}_{n,i}^{-1} = 0$ for all $i = 1, \ldots, p$. One can also check the second-order condition to verify that the solution is indeed a minimum. ■

B.2 Proof of Proposition 4.1

Let $\widetilde{\Delta}_n =: \mathsf{Diag}\big(\widetilde{\delta}_{n,1}, \ldots, \widetilde{\delta}_{n,p}\big)$. Then

$$\begin{aligned}
\mathcal{L}_n^{\mathrm{IS}}\left(\Sigma_n, U_n\widetilde{\Delta}_n U_n'\right) &= \frac{1}{p}\mathsf{Tr}\Big(\Sigma_n U_n\widetilde{\Delta}_n^{-1}U_n'\Big) - \frac{1}{p}\log\,\det\Big(\Sigma_n U_n\widetilde{\Delta}_n^{-1}U_n'\Big) - 1 \\
&= \frac{1}{p}\sum_{i=1}^{p}\Big[u_{n,i}'\Sigma_n u_{n,i}\cdot\widetilde{\delta}_{n,i}^{-1} + \log\big(\widetilde{\delta}_{n,i}\big)\Big] + \text{constant}\ .
\end{aligned}$$

The first-order condition is $-u_{n,i}'\Sigma_n u_{n,i}\cdot\widetilde{\delta}_{n,i}^{-2} + \widetilde{\delta}_{n,i}^{-1} = 0$ for all $i = 1, \ldots, p$. One can also check the second-order condition to verify that the solution is indeed a minimum. ■

B.3 Proof of Proposition 4.2

Let $\widetilde{\Delta}_n =: \mathsf{Diag}\big(\widetilde{\delta}_{n,1}, \ldots, \widetilde{\delta}_{n,p}\big)$. Then

$$\begin{aligned}
\mathcal{L}_n^{\mathrm{FR}}\left(\Sigma_n, U_n\widetilde{\Delta}_n U_n'\right) &= \frac{1}{p}\mathsf{Tr}\left[\Big(\Sigma_n - U_n\widetilde{\Delta}_n U_n'\Big)^2\right] = \frac{1}{p}\mathsf{Tr}\left[\Big(U_n'\Sigma_n U_n - \widetilde{\Delta}_n\Big)^2\right] \\
&= \frac{1}{p}\sum_{i=1}^{p}\Big(u_{n,i}'\Sigma_n u_{n,i} - \widetilde{\delta}_{n,i}\Big)^2\ ,
\end{aligned}$$

which is minimized if and only if $\widetilde{\delta}_{n,i} = u_{n,i}'\Sigma_n u_{n,i}$. ■

B.4 Proof of Proposition 4.3

Minimizing $\mathcal{L}^{\mathrm{MV}}$ is equivalent to minimizing

$$\frac{\mathsf{Tr}\left(\widetilde{\Sigma}_n^{-1}\Sigma_n\widetilde{\Sigma}_n^{-1}\right)}{\Big[\mathsf{Tr}\big(\widetilde{\Sigma}_n^{-1}\big)\Big]^2} = \frac{\mathsf{Tr}\left(U_n\widetilde{\Delta}_n^{-1}U_n'\Sigma_n U_n\widetilde{\Delta}_n^{-1}U_n'\right)}{\Big[\mathsf{Tr}\big(U_n\widetilde{\Delta}_n^{-1}U_n'\big)\Big]^2} = \frac{\mathsf{Tr}\left(\widetilde{\Delta}_n^{-1}U_n'\Sigma_n U_n\widetilde{\Delta}_n^{-1}\right)}{\Big[\mathsf{Tr}\big(\widetilde{\Delta}_n^{-1}\big)\Big]^2} \tag{B.1}$$

$$= \frac{\sum_{i=1}^{p}\widetilde{\delta}_{n,i}^{-2}\,\overline{d}_{n,i}}{\Big(\sum_{i=1}^{p}\widetilde{\delta}_{n,i}^{-1}\Big)^2} =: g(\widetilde{\boldsymbol{\delta}}_n) \tag{B.2}$$

The partial derivatives are, for all $i = 1, \ldots, p$,

$$\frac{\partial g}{\partial \widetilde{\delta}_{n,i}}\left(\widetilde{\boldsymbol{\delta}}_n\right) = \frac{-2\widetilde{\delta}_{n,i}^{-3}\,\overline{d}_{n,i}\left(\sum_{j=1}^{p}\widetilde{\delta}_{n,j}^{-1}\right)^2 + 2\widetilde{\delta}_{n,i}^{-2}\left(\sum_{j=1}^{p}\widetilde{\delta}_{n,j}^{-1}\right)\left(\sum_{j=1}^{p}\widetilde{\delta}_{n,j}^{-2}\,\overline{d}_{n,j}\right)}{\left(\sum_{j=1}^{p}\widetilde{\delta}_{n,j}^{-1}\right)^4} \tag{B.3}$$

It is only possible for all of them to be equal to zero if the ratio $\left(2\widetilde{\delta}_{n,i}^{-3}\,\overline{d}_{n,i}\right)/\left(2\widetilde{\delta}_{n,i}^{-2}\right)$ is equal to a constant independent of $i = 1, \ldots, p$. This means that the $\widetilde{\delta}_{n,i}$'s are proportional to the $\overline{d}_{n,i}$'s. We take the proportionality constant equal to one without loss of generality. One can also check the second-order condition to verify that the solution is indeed a minimum. ■

B.5 Proof of Proposition 4.4

As per Equation (1.7) of Gabor (1946), moving from a signal to its conjugate can be accomplished in either one of two mathematically equivalent ways: 1) taking the Fourier transform of the signal, suppressing the amplitudes belonging to negative frequencies, and multiplying the amplitudes of positive frequencies by two; or 2) taking the Hilbert transform of the signal. Appendix C.1 below gives a refresher course on the Hilbert transform. Then Appendix C.2 recalls that the Hilbert transform of the Cauchy probability density function

$$k^C(x) \coloneqq \frac{1}{\pi(x^2+1)} \qquad \text{is} \qquad \mathcal{H}_{k^C}(x) = -\frac{x}{\pi(x^2+1)}\,. \tag{B.4}$$

These expressions enable us to rewrite the smoothed Stein shrinker as the linear combination

$$\forall x \in \mathbb{R} \qquad \widehat{\theta}_n(x) = \frac{1}{p}\sum_{j=1}^{p}\frac{\pi}{h_n}\mathcal{H}_{k^C}\left(\frac{x-\lambda_{n,j}^{-1}}{h_n\lambda_{n,j}^{-1}}\right)\,. \tag{B.5}$$

Given that the operator that maps a signal into its Hilbert transform is both linear and anti-involutive (meaning that $\mathcal{H}_{\mathcal{H}_{k^C}} = -k^C$), applying Equation (B.4) to the smoothed Stein shrinker $\widehat{\theta}_n(x)$ yields its conjugate $\widehat{\theta}_n^*(x)$, and its squared amplitude $\mathcal{A}^2_{\widehat{\theta}_n}(x)$ follows. ■

C Some Foundations for Large-Dimensional Asymptotics

C.1 Hilbert Transform

To get started, we need to briefly recall a well-known important mathematical tool called the "Hilbert transform". Krantz (2009, p. 17) states that: "The Hilbert transform is, without question, the most important operator in analysis. It arises in many different contexts, and all these contexts are intertwined in profound and influential ways." It is defined as convolution with the "Cauchy kernel" $(\pi t)^{-1}$.

Definition C.1. *The Hilbert transform of a real function g is defined as*

$$\forall x \in \mathbb{R} \qquad \mathcal{H}_g(x) \coloneqq \frac{1}{\pi}PV\int_{-\infty}^{+\infty} g(t)\frac{dt}{t-x}\,. \tag{C.1}$$

PV represents the "Cauchy principal value", which is used to evaluate the singular integral in the following way:

$$PV\int_{-\infty}^{+\infty} g(t)\frac{dt}{t-x} \coloneqq \lim_{\varepsilon\to 0^+}\left[\int_{-\infty}^{x-\varepsilon} g(t)\frac{dt}{t-x} + \int_{x+\varepsilon}^{+\infty} g(t)\frac{dt}{t-x}\right] , \tag{C.2}$$

should this limit exist; otherwise, the Hilbert transform is not defined.

Recourse to the Cauchy principal value is needed because the Cauchy convolution kernel is singular, as a consequence of which the integral does not converge in the usual sense. The Hilbert transform is an anti-involution: $\mathcal{H}_{\mathcal{H}_g} = -g$; for example, see Titchmarsh (1948). Thus, g and $\mathcal{H}_g$ can be said to constitute a "Hilbert pair". Given that Eq. (1.7) of Gabor (1946) states

$$\mathcal{A}_g(x)^2 = g(x)^2 + \mathcal{H}_g(x)^2 , \tag{C.3}$$

from now on we will deal with the Hilbert transform instead of the amplitude in the proofs.

C.2 Cauchy Density

The Cauchy cumulative distribution function (c.d.f.) is defined as

$$\forall x \in \mathbb{R} \qquad K^C(x) \coloneqq \frac{1}{\pi}\arctan(x) + \frac{1}{2} . \tag{C.4}$$

Its corresponding probability density function (p.d.f.)

$$\forall x \in \mathbb{R} \qquad k^C(x) \coloneqq \frac{1}{\pi(x^2+1)} . \tag{C.5}$$

admits a well-known Hilbert transform; for example, see Poularikas (1998, Table 15.2):

$$\forall x \in \mathbb{R} \qquad \mathcal{H}_{k^C}(x) = -\frac{x}{\pi(x^2+1)} , \tag{C.6}$$

and their relationship is illustrated in Figure 9. Of all the known Hilbert pairs, the one with the Cauchy probability density function is the simplest, and this is no accident because the Hilbert transform is founded on both the Cauchy integration kernel and the Cauchy principal value.

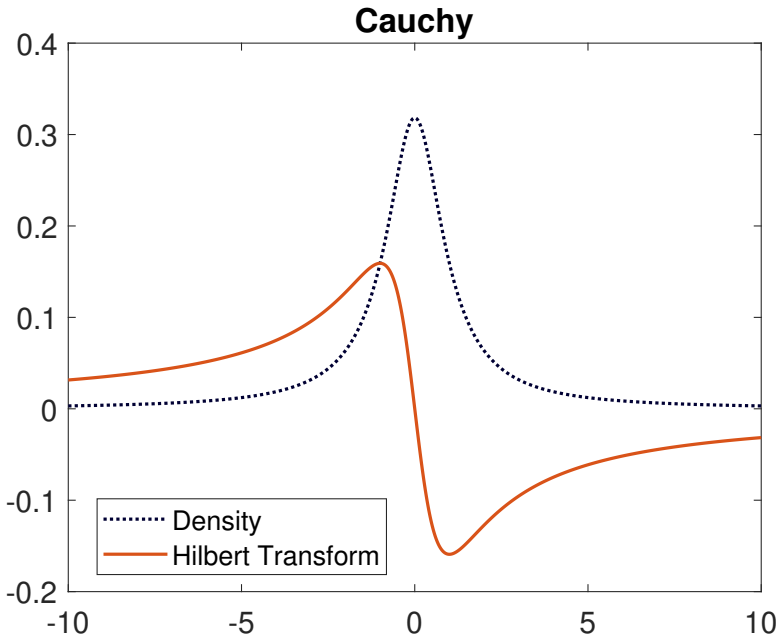

Figure 9: Comparison of the Cauchy density with its Hilbert transform.

The intuition behind the Hilbert transform is that it operates like a local attraction force. It is very positive if there are heavy mass points slightly larger than you, so it pushes you up (towards them), but very negative if they are slightly smaller, so it pushes you down (*also* towards them). When the mass points lie far away, it fades out to zero like gravitational attraction does. These effects are clearly apparent in Figure 9.

C.3 Stieltjes Transform

A transform closely related to the Hilbert transform is the "Stieltjes transform", which is defined on $\mathbb{C}^+$, the strict upper half of the complex plane. Given any bounded, nondecreasing function G, its Stieltjes transform m_G is defined as

$$\forall z \in \mathbb{C}^+ \qquad m_G(z) \coloneqq \int_{-\infty}^{+\infty} \frac{1}{x - z} dG(x) \ .$$

When G is sufficiently regular, including the existence of its derivative G', its Stieltjes transform admits an extension to the real line, which we denote as

$$\breve{m}_G(x) \coloneqq \lim_{z \in \mathbb{C}^+ \to x} m_G(z) \qquad \text{for all } x \in \mathbb{R} \ .$$

Note that, although $\breve{m}_G$ is a function of real argument, it is generally complex-valued. Its real and imaginary parts are given in terms of the derivative G' by, respectively,

$$\forall x \in \mathbb{R} \qquad \mathsf{Re}\big[\breve{m}_G(x)\big] = \pi \mathcal{H}_{G'}(x) \qquad \text{and} \qquad \mathsf{Im}\big[\breve{m}_G(x)\big] = \pi G'(x) \ . \tag{C.7}$$

Thus, any statement about the extension to the real line of the Stieltjes transform of a function is really a statement about the function's derivative and its Hilbert transform.

C.4 Spectral Distribution of the Precision Matrix

By analogy with the e.d.f. of the sample covariance matrix eigenvalues, F_n, we can construct the e.d.f. of the sample precision matrix eigenvalues:

$$\forall x \in \mathbb{R} \qquad \Phi_n(x) \coloneqq \frac{1}{p} \sum_{i=1}^{p} \mathbf{1}_{\{\lambda_{n,i}^{-1} \leq x\}} \ . \tag{C.8}$$

Note that the smallest sample covariance matrix eigenvalue $\lambda_{n,1}$ is strictly positive with probability one under Assumptions 3.1–3.2; therefore, the relationship between Φ_n and F_n can be expressed symmetrically as follows:

$$\forall x \in \mathbb{R} \qquad \Phi_n(x) = \begin{cases} 1 - F_n(1/x) & \text{if } x > 0 \\ 0 & \text{otherwise} \end{cases} \tag{C.9}$$

$$\forall x \in \mathbb{R} \qquad F_n(x) = \begin{cases} 1 - \Phi_n(1/x) & \text{if } x > 0 \\ 0 & \text{otherwise} \ . \end{cases} \tag{C.10}$$

Now define

$$\forall x \in \mathbb{R} \qquad \Phi(x) \coloneqq \begin{cases} 1 - F(1/x) & \text{if } x > 0 \\ 0 & \text{otherwise} \end{cases} \ , \tag{C.11}$$

where $F(x)$ is the limiting spectral distribution of the sample eigenvalues. The mapping in (C.9) implies that, under Assumptions 3.1–3.2, large-dimensional asymptotics in precision-matrix space shares all the nice properties of large-dimensional asymptotics in covariance-matrix space proven by Silverstein and Bai (1995), Silverstein (1995), Silverstein and Choi (1995), and Bai and Silverstein (1998), namely:

1. $\forall x \in \mathbb{R}, \ \ \Phi_n(x) \to \Phi(x)$ almost surely.
2. The limiting precision spectral distribution Φ has a continuous derivative ϕ on $\mathbb{R}$.
3. The limiting precision spectral density $\phi(x)$ has a Hilbert transform $\mathcal{H}_\phi(x)$ that also exists and is continuous on $\mathbb{R}$.
4. As a consequence, the Stieltjes transform of Φ admits a complex-valued extension to the real line $\breve{m}_\Phi$ that also exists and is continuous on $\mathbb{R}$.
5. $\mathsf{Supp}(\Phi)$ is the union of a finite number $\nu \geq 1$ of compact intervals: $\mathsf{Supp}(\Phi) = \bigcup_{k=1}^{\nu} [1/b_k, 1/a_k]$, where $0 < a_1 < b_1 < \cdots < a_\nu < b_\nu < \infty$.
6. Φ is uniquely determined by c and H, so we can denote it more explicitly by $\Phi_{c,H}$ whenever there is some risk of ambiguity.

So we can work with Φ_n, Φ, ϕ, $\mathcal{H}_\phi$, and $\breve{m}_\Phi$ just like with F_n, F, f, $\mathcal{H}_f$, and $\breve{m}_F$.

C.5 First Incomplete Moment Function

At times, it can be mathematically convenient to work not with a cumulative distribution function but with the first incomplete moment function associated with it. Given any generic

c.d.f G, its first incomplete moment function is defined as

$$\forall x \in \mathbb{R} \qquad LG(x) := \int_{-\infty}^{x} t \, dG(t) \ . \tag{C.12}$$

The prefix L is meant to represent the fact that the mapping from a c.d.f. to its first incomplete moment function is *L*inear. The Stieltjes transform of the first incomplete moment function can be easily deduced from that of its corresponding c.d.f. as follows:

$$\forall x \in \mathbb{R} \qquad \breve{m}_{LG}(x) = 1 + x\breve{m}_G(x) \ . \tag{C.13}$$

Of particular interest is applying the L transform to the limiting spectral distibution F. Intuitively, this means giving more importance to the big eigenvalues, relative to smaller ones. This can be seen most clearly when all population eigenvalues are equal to one, since a closed-form solution for the (limiting) density of the sample eigenvalues was found by Marčenko and Pastur (1967). Figure 10 illustrates visually.

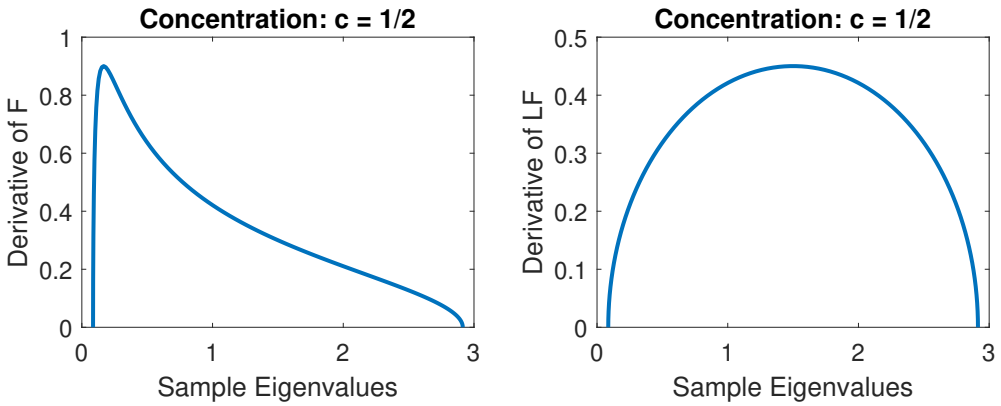

Figure 10: Comparison of F with LF.

From the density of the sample eigenvalues on the left-hand side, one can see that there are many small eigenvalues and few large ones, which gives the graph a lopsided appearance. By contrast, on the right-hand side, overweighting the large eigenvalues through the L operator means that these few large eigenvalues take on more importance, which yields an elegantly symmetrical graph: the semi-circular law made famous by Wigner (1955).

C.6 Initial Findings About the Optimal Modulating Function

A key mathematical object is obtained by combining Sections C.5 and C.4, meaning we construct the first incomplete moment of precision matrix eigenvalues:

$$\forall x \in \mathbb{R} \qquad \Psi_n(x) := L\Phi_n(x) := \frac{1}{p} \sum_{i=1}^{p} \lambda_{n,i}^{-1} \, \mathbf{1}_{\{\lambda_{n,i}^{-1} \leq x\}} \ . \tag{C.14}$$

Needless to say, the L operator preserves all of the nice properties 1–6 enumerated in Section C.4; therefore, it is as easy to work with Ψ_n, $\Psi := L\Phi$, $\psi := \Psi'$, $\mathcal{H}_\psi$, and $\breve{m}_\Psi$ as with Φ_n, Φ, ϕ, $\mathcal{H}_\phi$, and $\breve{m}_\Phi$; or, for that matter, as with F_n, F, f, $\mathcal{H}_f$, and $\breve{m}_F$ in the first place.

1. $\forall x \in \mathbb{R}, \;\; \Psi_n(x) \to \Psi(x)$ almost surely.
2. The limiting precision spectral distribution Ψ has a continuous derivative ψ on $\mathbb{R}$.
3. The limiting precision spectral density $\psi(x)$ has a Hilbert transform $\mathcal{H}_\psi(x)$ that also exists and is continuous on $\mathbb{R}$.
4. As a consequence, the Stieltjes transform of Ψ admits a complex-valued extension to the real line $\breve{m}_\Psi$ that also exists and is continuous on $\mathbb{R}$.
5. $\mathsf{Supp}(\Psi)$ is the union of a finite number $\nu \geq 1$ of compact intervals: $\mathsf{Supp}(\Psi) = \bigcup_{k=1}^{\nu}[1/b_k, 1/a_k]$, where $0 < a_1 < b_1 < \cdots < a_\nu < b_\nu < \infty$.
6. Ψ is uniquely determined by c and H, so we can denote it more explicitly by $\Psi_{c,H}$ whenever there is some risk of ambiguity.

As a result, all the asymptotic results that have been obtained for kernel estimation of the limiting sample spectral density f by Jing et al. (2010), and subsequently extended to kernel estimation of its Hilbert transform $\mathcal{H}_f$ by Ledoit and Wolf (2020), will carry through to $\mathcal{H}_\psi$. This is important because the modulating function featured in Proposition 2.1's reinterpretation of Stein's (1986) formula as "Linear-Inverse Shrinkage" is merely a trivial rescaling of the function $\mathcal{H}_{\Psi_n'}(x)$, as the next proposition shows.

Proposition C.1.

$$\forall x \notin \{\lambda_{n,1}, \ldots, \lambda_{n,p}\} \qquad \widetilde{\theta}_n(x) = \frac{\pi\, p}{p-1}\, \mathcal{H}_{\Psi_n'}(x) \;. \tag{C.15}$$

Proof of Proposition C.1. From (C.1), the Hilbert transform of a generic p.d.f. g with corresponding c.d.f. G is

$$\mathcal{H}_g(x) \coloneqq \frac{1}{\pi} PV \int_{-\infty}^{+\infty} \frac{1}{t-x}\, dG(t) \;. \tag{C.16}$$

Plugging Equation (C.14) into Equation (C.16) yields

$$\forall x \notin \{\lambda_{n,1}, \ldots, \lambda_{n,p}\} \qquad \mathcal{H}_{\Psi_n'}(x) = \frac{1}{\pi} PV \int_{-\infty}^{+\infty} \frac{1}{t-x}\, d\Psi_n(t) \tag{C.17}$$

$$= \frac{1}{\pi} \cdot \frac{1}{p} \sum_{j=1}^{p} \frac{\lambda_{n,j}^{-1}}{\lambda_{n,j}^{-1} - x} \tag{C.18}$$

$$= \frac{p-1}{\pi\, p}\, \widetilde{\theta}_n(x) \;. \blacksquare \tag{C.19}$$

To unpack Proposition C.1 intuitively: (i) the Hilbert transform acts as a local attraction force; (ii) the L operator acknowledges that eigenvalues are influential in proportion to their magnitudes; and (iii) Stein's formula linearly shrinks precision matrix eigenvalues, hence $(L\Phi_n)'$ instead of $(LF_n)'$.

It is impossible to work directly with $\widetilde{\theta}_n(x)$ or $\mathcal{H}_{\Psi_n'}(x)$ because they both explode when x is equal to the inverse of a sample eigenvalue. This is why Stein (1975) had so much trouble with his nonlinear shrinkage formula that he had to post-process it through a numerical regularization process called isotonization, described in detail in the Appendix of Lin and Perlman (1985).

The more fruitful approach is to take an indirect route that starts with the much better-behaved function $\Psi_n(x)$. To this end, we establish the following lemma.

Lemma C.1. *Under Assumptions 3.1–3.2,*

$$\forall x \in \mathbb{R} \qquad \Psi_n(x) \xrightarrow{a.s.} \Psi(x) \coloneqq L\Phi(x) \ ; \tag{C.20}$$

furthermore, the function $\Psi(x)$ admits a continuous derivative $\psi(x)$ that has a well-defined Hilbert transform $\mathcal{H}_\psi(x)$ on all of $x \in \mathbb{R}$.

Proof of Lemma C.1. Silverstein (1995, Theorem 1.1) proves that

$$F_n(x) \xrightarrow{\text{a.s.}} F(x) \tag{C.21}$$

under assumptions that are even less restrictive than Assumptions 3.1–3.2. He proves it only for x where the limiting spectral c.d.f. F is continuous. However, given that we assume that the limiting concentration ratio $c = \lim_{n\to\infty} p/n$ is strictly below 1, and that the support of the distribution of population covariance matrix eigenvalues $\mathsf{Supp}(H)$ is bounded away from zero, the results of Silverstein and Choi (1995) imply that F is continuous on $\mathbb{R}$, hence (C.21) holds for all $x \in \mathbb{R}$.

Next, injecting Equations (C.9) and (C.11) into (C.21) implies that

$$\forall x \in \mathbb{R} \qquad \Phi_n(x) \xrightarrow{\text{a.s.}} \Phi(x) \ . \tag{C.22}$$

There might be some concern about what happens near zero, given that we are inverting the sample eigenvalues. However, this is a moot point because with probability one there will be no eigenvalues in $(-\infty, a_1/2]$ for all n sufficiently large (Bai and Silverstein, 1998). Here $a_1 > 0$ denotes the lower bound of the support of F (cf. Section 3.1).

Finally, moving from the c.d.f. to the first incomplete moment function as per Section (C.5) is 'for free' because the L-transform is linear; therefore:

$$\forall x \in \mathbb{R} \qquad \Psi_n(x) = L\Phi_n(x) \xrightarrow{\text{a.s.}} \Psi(x) \coloneqq L\Phi(x) \ . \tag{C.23}$$

The function $\Psi(x)$ admits a continuous derivative $\psi(x)$ that has a well-defined Hilbert transform $\mathcal{H}_\psi(x)$ on all of $x \in \mathbb{R}$ because F does, and the successive transformations preserve these properties, given that the support of F is bounded away from zero. ■

The next order of business is to show that ψ (respectively, its Hilbert transform $\mathcal{H}_\psi$) is consistently estimated by convolving the first incomplete moment function of precision matrix eigenvalues Ψ_n with the Cauchy density (respectively, its Hilbert transform), provided that the kernel bandwidth parameter is chosen within an appropriate range.

D Cauchy Kernel Estimation in Inverse-Eigenvalues Space

At this juncture, we turn to a well-known technique called kernel estimation. In the related context of estimating the limiting sample spectral density f, this technique has already been used

by Jing et al. (2010). There are several differences between our work and theirs. First, we apply kernel estimation to a bigger ultimate problem, namely, the estimation of the covariance matrix as a whole, as opposed to the estimation of the limiting e.d.f. of sample eigenvalues. Second, we employ a different kernel, namely, the Cauchy density, as opposed to the Gaussian density. Third, we use a locally adaptive bandwidth, whereas they use a globally uniform bandwidth. Fourth, we apply kernel estimation at the level of the first incomplete moment function rather than the cumulative distribution function. And fifth, we also estimate the Hilbert transform.

D.1 Formulas for Kernel Estimators

From here on, for notational simplicity, all the proofs assume that the support of F, denoted by $\mathsf{Supp}(F)$, is a finite interval $[a, b]$, where $0 < a < b < \infty$. At the cost of increased notational complexity, all the proofs also extend to the general case where $\mathsf{Supp}(F)$ is the union of a finite number $\nu \geq 1$ of compact intervals: $\mathsf{Supp}(F) = \bigcup_{k=1}^{\nu}[a_k, b_k]$, where $0 < a_1 < b_1 < \cdots < a_\nu < b_\nu < \infty$.

Remark D.1. For the purpose of clarity, note that the Cauchy density that we use in our nonparametric kernel estimator should not be confused with the Cauchy convolution kernel that defines the Hilbert transform, because they are not the same — although there is a deep mathematical relationship between the two concepts. Thus, we will try to stay away from the unqualified expression 'Cauchy kernel", which can be slightly ambiguous. ■

If we adopt a locally adaptive (proportional) bandwidth $h_{n,j} := h_n \lambda_{n,j}^{-1}$, for $j = 1, \ldots, p$, in conjunction with a generic kernel k, the kernel estimators of the spectral density of the population covariance matrix ϕ and its Hilbert transform are given by:

$$\forall x \in \mathbb{R} \qquad \widehat{\phi}_n(x) := \frac{1}{p}\sum_{j=1}^{p} \frac{1}{h_{n,j}} \cdot k\left(\frac{x - \lambda_{n,j}^{-1}}{h_{n,j}}\right) \tag{D.1}$$

$$\forall x \in \mathbb{R} \qquad \mathcal{H}_{\widehat{\phi}_n}(x) := \frac{1}{p}\sum_{j=1}^{p} \frac{1}{h_{n,j}} \cdot \mathcal{H}_k\left(\frac{x - \lambda_{n,j}^{-1}}{h_{n,j}}\right) \tag{D.2}$$

Given formula (D.1), the kernel estimator of the spectral c.d.f. is given by

$$\widehat{\Phi}_n(x) = \int_{\infty}^{x} \widehat{\phi}_n(t)dt = \frac{1}{p}\sum_{j=1}^{p} \frac{1}{h_{n,j}} \cdot K\left(\frac{x - \lambda_{n,j}^{-1}}{h_{n,j}}\right) , \tag{D.3}$$

with $K(x) := \int_{-\infty}^{x} k(t)dt$. Analogously, the kernel estimators of the function ψ and its Hilbert transform are given by:

$$\forall x \in \mathbb{R} \qquad \widehat{\psi}_n(x) := \frac{1}{p}\sum_{j=1}^{p} \frac{\lambda_{n,j}^{-1}}{h_{n,j}} \cdot k\left(\frac{x - \lambda_{n,j}^{-1}}{h_{n,j}}\right) \tag{D.4}$$

$$\forall x \in \mathbb{R} \qquad \mathcal{H}_{\widehat{\psi}_n}(x) := \frac{1}{p}\sum_{j=1}^{p} \frac{\lambda_{n,j}^{-1}}{h_{n,j}} \cdot \mathcal{H}_k\left(\frac{x - \lambda_{n,j}^{-1}}{h_{n,j}}\right) \tag{D.5}$$

Equations (C.5)–(C.6) then imply that for the special choice of the Cauchy density for the kernel k, the kernel estimators of the function ψ and its Hilbert transform specialize to:

$$\forall x \in \mathbb{R} \qquad \widehat{\psi}_n(x) = \frac{1}{p}\sum_{j=1}^{p} \frac{\lambda_{n,j}^{-1}}{h_{n,j}} \cdot \frac{1}{\pi\left[\left(\dfrac{x-\lambda_{n,j}^{-1}}{h_{n,j}}\right)^2 + 1\right]} \tag{D.6}$$

$$= \frac{1}{p}\sum_{j=1}^{p} \lambda_{n,j}^{-1} \frac{h_{n,j}}{\pi\left[\left(x-\lambda_{n,j}^{-1}\right)^2 + h_{n,j}^2\right]} \tag{D.7}$$

$$= \frac{1}{p}\sum_{j=1}^{p} \lambda_{n,j}^{-1} \frac{h_n\lambda_{n,j}^{-1}}{\pi\left[\left(x-\lambda_{n,j}^{-1}\right)^2 + h_n^2\lambda_{n,j}^{-2}\right]} \tag{D.8}$$

$$\forall x \in \mathbb{R} \qquad \mathcal{H}_{\widehat{\psi}_n}(x) = -\frac{1}{p}\sum_{j=1}^{p} \frac{\lambda_{n,j}^{-1}}{h_{n,j}} \cdot \frac{\dfrac{x-\lambda_{n,j}^{-1}}{h_{n,j}}}{\pi\left[\left(\dfrac{x-\lambda_{n,j}^{-1}}{h_{n,j}}\right)^2 + 1\right]} \tag{D.9}$$

$$= -\frac{1}{p}\sum_{j=1}^{p} \lambda_{n,j}^{-1} \frac{x-\lambda_{n,j}^{-1}}{\pi\left[\left(x-\lambda_{n,j}^{-1}\right)^2 + h_{n,j}^2\right]} \tag{D.10}$$

$$= -\frac{1}{p}\sum_{j=1}^{p} \lambda_{n,j}^{-1} \frac{x-\lambda_{n,j}^{-1}}{\pi\left[\left(x-\lambda_{n,j}^{-1}\right)^2 + h_n^2\lambda_{n,j}^{-2}\right]} \tag{D.11}$$

Comparing Equation (D.11) with Equation (3.2) reveals that

$$\forall x \in \mathbb{R} \qquad \widehat{\theta_n}(x) = \pi\mathcal{H}_{\widehat{\psi}_n}(x) \ . \tag{D.12}$$

D.2 Technical Preliminaries Regarding the Kernel

Assumption D.1 (Kernel)**.** *Let k denote a continuously differentiable, symmetric, real-valued, probability density function (p.d.f.), whose c.d.f. is denoted by K, that satisfies the following conditions:*

a. $k(x) = o(1/x)$.

b. Its Hilbert transform $\mathcal{H}_k$ exists and is continuous on $\mathbb{R}$.

c. Both the kernel k and its Hilbert transform $\mathcal{H}_k$ are functions of bounded variation.

d. $\lim_{x\to+\infty} x\mathcal{H}_k(x) = -1/\pi$;

Note that Assumption D.1.c implies that

$$\int_{-\infty}^{+\infty} \left|\frac{d\breve{m}_K}{dx}(x)\right| dx < \infty \ , \ . \tag{D.13}$$

Indeed, the imaginary part of (D.13) is taken care of, since the imaginary part of $\breve{m}_K$ is π times the kernel density k itself, which is assumed to be of bounded variation. The real part of (D.13)

is also taken care of, since the real part of $\breve{m}_K$ is π times the Hilbert transform of the kernel density k, which is also assumed to be of bounded variation. From now on, we define

$$\breve{m}'_K(x) \coloneqq \frac{d\breve{m}_K}{dx}(x) \ .$$

As for Assumption D.1.d, it essentially means that the Hilbert transform decays away from the bulk of the kernel density at the same speed as for compactly-supported kernels. The point is to rule out kernels with pathological tail behavior. The presence of the coefficient $-1/\pi$ is solely due to the conventions we used in defining the Hilbert transform in (C.1). Together, Assumptions D.1.b and D.1.d imply that

$$\exists C > 0 \quad \text{s.t.} \quad \forall x \in \mathbb{R} \quad \left| x\mathcal{H}_k(x) \right| \leq C < \infty \ . \tag{D.14}$$

Lemma D.1. *If a kernel k satisfies Assumption D.1 then*

$$\lim_{x \to +\infty} \frac{1}{x} \int_{-x}^{x} |\breve{m}_K(t)| dt = 0 \ . \tag{D.15}$$

Proof of Lemma D.1. Fix any $\varepsilon \in (0, 1/\pi)$. By Assumption D.1.a, there exists $x_1 > 0$ such that $\forall x > x_1 \ \ k(x) < \varepsilon/x$. By Assumption D.1.d, there exists $x_2 > x_1$ such that

$$\forall x > x_2 \qquad \left(-\frac{1}{\pi} - \varepsilon \right) \frac{1}{x} < \mathcal{H}_k(x) < \left(-\frac{1}{\pi} + \varepsilon \right) \frac{1}{x} < 0 \ .$$

Therefore, for all $x > x_2$, $|\breve{m}_K(x)| = |\pi \mathcal{H}_k(x) + i\pi k(x)| < (1 + 2\pi\varepsilon)/x$. Recall also that both k and $\mathcal{H}_k$ are functions of bounded variation, so $|\breve{m}_K(x)|$ is bounded over $x \in \mathbb{R}$ by a finite upper bound, which we can call C_1. Then we have

$$\begin{aligned} \forall x > x_2 \qquad \frac{1}{x} \int_{-x}^{x} |\breve{m}_K(t)| dt &= \frac{1}{x} \int_{-x_2}^{x_2} |\breve{m}_K(t)| dt + \frac{2}{x} \int_{x_2}^{x} |\breve{m}_K(t)| dt \\ &\leq \frac{2x_2 C_1}{x} + \frac{2}{x} \int_{x_2}^{x} \frac{1 + 2\pi\varepsilon}{t} dt \\ &\leq \frac{2x_2 C_1}{x} + 2(1 + 2\pi\varepsilon) \frac{\log(x) - \log(x_2)}{x} \ , \end{aligned}$$

which converges to zero as $x \to +\infty$. ■

Lemma D.2. *The Cauchy density satisfies Assumption D.1.*

Proof of Lemma D.2. Follows directly from visual inspection of Equations (C.5)–(C.6). ■

Assumption D.2. *The proportional bandwidths are of the form $h_{n,j} \coloneqq h_n \lambda_{n,j}^{-1}$, for $j = 1, \ldots, p$, where h_n is a sequence of positive numbers satisfying*

$$\lim_{n \to \infty} n h_n^{5/2} = \infty \qquad \textit{and} \qquad \lim_{n \to \infty} h_n = 0 \ . \tag{D.16}$$

Lemma D.3. *For a choice of bandwidth h_n satisfying Assumption D.2 and a kernel k satisfying Assumption D.1,*

$$\lim_{n \to \infty} PV \int \frac{1}{1 + u h_n} k(u) du = 1 \ . \tag{D.17}$$

Proof of Lemma D.3.

$$PV\int \frac{1}{1+uh_n}k(u)du = \frac{1}{h_n}PV\int \frac{1}{u-(-1/h_n)}k(u)du$$
$$= \frac{\pi}{h_n}\mathcal{H}_k\left(-\frac{1}{h_n}\right) = -\frac{\pi}{h_n}\mathcal{H}_k\left(\frac{1}{h_n}\right) \ .$$

The desired result follows from injecting $x := 1/h_n$ into Assumption D.1.d and letting both n and x go to infinity. ■

D.3 Some Useful Lemmas

This section lays the groundwork for the main proofs. It is a mix between extending results in the existing literature from F to Ψ and developing newer material.

Lemma D.4. *For all $x \in \mathbb{R}$, $x \neq 0$,*

$$\lim_{z\in\mathbb{C}^+\to x} m_\Psi(z) =: \breve{m}_\Psi(x) \quad \textit{exists.} \tag{D.18}$$

The function $\breve{m}_\Psi$ is continuous on $\mathbb{R}\setminus\{0\}$. Consequently, Ψ has a continuous derivative ψ on $\mathbb{R}\setminus\{0\}$ given by $\psi(x) = \frac{1}{\pi}\mathsf{Im}\left[\breve{m}_\Psi(x)\right]$.

Proof of Lemma D.4.

$$\forall z\in\mathbb{C}^+ \qquad \breve{m}_\Phi(z) = \int_0^{+\infty}\frac{1}{t-z}d\Phi(t) = \int_0^{+\infty}\frac{1}{t-z}\cdot\frac{1}{t^2}dF(1/t) \tag{D.19}$$
$$= \int_{+\infty}^0 \frac{1}{\frac{1}{u}-z}\cdot u^2\left(-\frac{1}{u^2}\right)dF(u) = \int_0^{+\infty}\frac{u}{1-uz}dF(u) \tag{D.20}$$
$$= \frac{1}{z}\int_0^{+\infty}\frac{u}{\frac{1}{z}-u}dF(u) = \frac{1}{z}\int_0^{+\infty}\frac{u-\frac{1}{z}+\frac{1}{z}}{\frac{1}{z}-u}dF(u) \tag{D.21}$$
$$= -\frac{1}{z}+\frac{1}{z^2}\int_0^{+\infty}\frac{1}{\frac{1}{z}-u}dF(u) = -\frac{1}{z}-\frac{1}{z^2}m_F\left(1/z\right) \tag{D.22}$$
$$m_\Psi(z) = 1+zm_\Phi(z) = -\frac{1}{z}m_F\left(1/z\right) \ . \tag{D.23}$$

The first part of Lemma D.4 follows from Equation (D.23) and from Theorem 1.1 of Silverstein and Choi (1995), which states that for all $x\in\mathbb{R}$, $x\neq 0$,

$$\lim_{z\in\mathbb{C}^+\to x} m_F(z) =: \breve{m}_F(x) \quad \text{exists,} \tag{D.24}$$

and the function $\breve{m}_F$ is continuous on $\mathbb{R}\setminus\{0\}$. The fact that Ψ has a continuous derivative ψ on $\mathbb{R}\setminus\{0\}$ given by $\psi(x) = \frac{1}{\pi}\mathsf{Im}\left[\breve{m}_\Psi(x)\right]$ then follows from applying Theorem 2.1 of Silverstein and Choi (1995) to the c.d.f. $\Psi(x)/\Psi(b')$. ■

Much of the work is carried out on an interval $[a',b']$ such that $0<a'<a$ and $b<b'<+\infty$. Due to Assumption 3.2.d, it is possible to choose a' and b' so that there exists N such that $\forall n\geq N,\ \ \mathsf{Supp}\left(F_{c_n,H_n}\right)\subsetneq[a',b']$. Thanks to Assumption 3.1, it is possible to assume without loss of generality that $\forall n\geq N,\ \ c_n\in[c/2,(c+1)/2]$.

Lemma D.5. *Under Assumptions 3.1-3.2, let $\Psi_{c_n,H_n}(x)$ be the c.d.f. obtained from $\Psi_{c,H}(x)$ by replacing c and H with c_n and H_n, respectively. Furthermore, let $\breve{m}_{\Psi_{c_n,H_n}}(x)$ denote the extension to the real line of the Stieltjes transform of $\Psi_{c_n,H_n}(x)$. Then there exists a fixed, finite upper bound M such that*

$$\sup_{n\geq N, x\in\mathbb{R}} \left|\breve{m}_{\Psi_{c_n,H_n}}(x)\right| \leq M \ . \tag{D.25}$$

Proof of Lemma D.5. The proof starts from the associated c.d.f.

$$\forall x \in \mathbb{R} \qquad \underline{F}_{c,H}(x) := cF_{c,H}(x) - (c-1)\mathbf{1}_{\{x\geq 0\}} \ , \tag{D.26}$$

It is the limiting spectral c.d.f. of the matrix Y_nY_n'/n, which has the same eigenvalues as the sample covariance matrix, apart from $|p-n|$ null eigenvalues. We define $\underline{F}_{c_n,H_n}$ in similar fashion. Equation (5.5) of Jing et al. (2010) implies that there exists a fixed, finite upper bound M_1 such that

$$\sup_{n\geq N, x\in[a',b']} \left|\breve{m}_{\underline{F}_{c_n,H_n}}(x)\right| \leq M_1 \ . \tag{D.27}$$

The relationship

$$\forall x \in (0,+\infty) \qquad \breve{m}_{F_{c_n,H_n}}(x) = \frac{1-c_n}{c_n x} + \frac{1}{c_n}\breve{m}_{\underline{F}_{c_n,H_n}}(x) \ , \tag{D.28}$$

which follows directly from Equation (1.3) of Silverstein (1995), guarantees that

$$\sup_{n\geq N, x\in[a',b']} \left|\breve{m}_{F_{c_n,H_n}}(x)\right| \leq \frac{1-c}{ca'} + \frac{2}{c}M_1 =: M_2 < \infty \ . \tag{D.29}$$

Next, we move into the realm of precision matrix eigenvalues: for all $x \in (0,+\infty)$,

$$\breve{m}_{\Phi_{c_n,H_n}}(x) = \int_0^{+\infty} \frac{1}{t-x} d\Phi_{c_n,H_n}(t) = \int_0^{+\infty} \frac{1}{t-x}\cdot\frac{1}{t^2} dF_{c_n,H_n}(1/t) \tag{D.30}$$

$$= \int_{+\infty}^0 \frac{1}{\frac{1}{u}-x}\cdot u^2\left(-\frac{1}{u^2}\right) dF_{c_n,H_n}(u) = \int_0^{+\infty} \frac{u}{1-ux} dF_{c_n,H_n}(u) \tag{D.31}$$

$$= \frac{1}{x}\int_0^{+\infty} \frac{u}{\frac{1}{x}-u} dF_{c_n,H_n}(u) = \frac{1}{x}\int_0^{+\infty} \frac{u-\frac{1}{x}+\frac{1}{x}}{\frac{1}{x}-u} dF_{c_n,H_n}(u) \tag{D.32}$$

$$= -\frac{1}{x} + \frac{1}{x^2}\int_0^{+\infty} \frac{1}{\frac{1}{x}-u} dF_{c_n,H_n}(u) \tag{D.33}$$

$$= -\frac{1}{x} - \frac{1}{x^2}\breve{m}_{F_{c_n,H_n}}(1/x) \ . \tag{D.34}$$

Then we turn to the first incomplete moment function. From Section C.5:

$$\forall n \in \mathbb{N} \quad \forall x \in (0,+\infty) \qquad \breve{m}_{\Psi_{c_n,H_n}}(x) = 1 + x\breve{m}_{\Phi_{c_n,H_n}}(x) = -\frac{1}{x}\breve{m}_{F_{c_n,H_n}}(1/x) \ , \tag{D.35}$$

from which we deduce

$$\sup_{n\geq N, x\in[1/b',1/a']} \left|\breve{m}_{\Psi_{c_n,H_n}}(x)\right| \leq b'M_2 =: M < \infty \ . \tag{D.36}$$

As for $x \notin [1/b', 1/a']$, $\mathsf{Im}\left[\breve{m}_{\Psi_{c_n,H_n}}(x)\right] = 0$ because $\forall n \geq N$, $\mathsf{Supp}\left(\Psi_{c_n,H_n}\right) \subsetneq [1/b', 1/a']$, so we need only worry about the real part:

$$\forall x \in (-\infty, 1/b') \quad 0 \leq \mathsf{Re}\left[\breve{m}_{\Psi_{c_n,H_n}}(x)\right] = \int_{1/b'}^{1/a'} \frac{1}{t-x}\, d\Psi_{c_n,H_n}(t) \tag{D.37}$$

$$\leq PV \int_{1/b'}^{1/a'} \frac{1}{t-\frac{1}{b'}} d\Psi_{c_n,H_n}(t) = \mathsf{Re}\left[\breve{m}_{\Psi_{c_n,H_n}}(1/b')\right]$$

$$\leq M < \infty \tag{D.38}$$

$$\forall x \in (1/a', +\infty) \quad 0 \geq \mathsf{Re}\left[\breve{m}_{\Psi_{c_n,H_n}}(x)\right] = \int_{1/b'}^{1/a'} \frac{1}{t-x}\, d\Psi_{c_n,H_n}(t) \tag{D.39}$$

$$\geq PV \int_{1/b'}^{1/a'} \frac{1}{t-\frac{1}{a'}} d\Psi_{c_n,H_n}(t) = \mathsf{Re}\left[\breve{m}_{\Psi_{c_n,H_n}}(1/a')\right]$$

$$\geq -M > -\infty\ . \tag{D.40}$$

This concludes the proof of Lemma D.5. ■

Remark D.2. The result also holds with $\Psi_{c,H}$ in place of Ψ_{c_n,H_n}. ■

Lemma D.6. *Under Assumptions 3.1-3.2, D.1, and D.2,*

$$\mathbb{E}\left\{\sup_{x\in\left[\frac{1}{b'},\frac{1}{a'}\right]}\left|\Psi_n(x) - \Psi_{c_n,H_n}(x)\right|\right\} = O\left(\frac{1}{n^{2/5}}\right)\ . \tag{D.41}$$

Proof of Lemma D.6. Theorem 3 of Jing et al. (2010) implies that, under Assumptions 3.1-3.2, D.1, and D.2,

$$\mathbb{E}\left\{\sup_{x\in[a',b']}\left|F_n(x) - F_{c_n,H_n}(x)\right|\right\} = O\left(\frac{1}{n^{2/5}}\right)\ . \tag{D.42}$$

Moving to precision matrix eigenvalues, we have:

$$\forall x \in [a', b'] \qquad \Phi_n(x) = 1 - F_n\left(\frac{1}{x}\right) \tag{D.43}$$

$$\Phi_{c_n,H_n}(x) = 1 - F_{c_n,H_n}\left(\frac{1}{x}\right) \tag{D.44}$$

$$\Phi_n(x) - \Phi_{c_n,H_n}(x) = F_{c_n,H_n}\left(\frac{1}{x}\right) - F_n\left(\frac{1}{x}\right) \tag{D.45}$$

$$\sup_{x\in\left[\frac{1}{b'},\frac{1}{a'}\right]}\left|\Phi_n(x) - \Phi_{c_n,H_n}(x)\right| = \sup_{x\in[a',b']}\left|F_n\left(x\right) - F_{c_n,H_n}\left(x\right)\right|\ . \tag{D.46}$$

This sets the stage for the last transformation:

$$\forall x \in \left[1/b', 1/a'\right] \qquad \Psi_n(x) = L\Phi_n(x) \tag{D.47}$$

$$\Psi_{c_n,H_n}(x) = L\Phi_{c_n,H_n}(x) \tag{D.48}$$

$$\Psi_n(x) - \Psi_{c_n,H_n}(x) = L\left[\Phi_n(x) - \Phi_{c_n,H_n}(x)\right] \tag{D.49}$$

$$= \int_{1/b'}^{x} t\left[d\Phi_n(t) - d\Phi_{c_n,H_n}(t)\right]\ . \tag{D.50}$$

Using integration by parts, we get, for all $x \in [1/b', 1/a']$,

$$\Psi_n(x) - \Psi_{c_n,H_n}(x) = x\big[\Phi_n(x) - \Phi_{c_n,H_n}(x)\big] - \int_{1/b'}^{x} \big[\Phi_n(t) - \Phi_{c_n,H_n}(t)\big]\, dt$$

$$\big|\Psi_n(x) - \Psi_{c_n,H_n}(x)\big| \leq \frac{1}{a'}\big|\Phi_n(x) - \Phi_{c_n,H_n}(x)\big| + \frac{1}{a'} \sup_{x \in \left[\frac{1}{b'}, \frac{1}{a'}\right]} \big|\Phi_n(x) - \Phi_{c_n,H_n}(x)\big|$$

$$\sup_{x \in \left[\frac{1}{b'}, \frac{1}{a'}\right]} \big|\Psi_n(x) - \Psi_{c_n,H_n}(x)\big| \leq \frac{2}{a'} \sup_{x \in \left[\frac{1}{b'}, \frac{1}{a'}\right]} \big|\Phi_n(x) - \Phi_{c_n,H_n}(x)\big| \tag{D.51}$$

By putting Equations (D.42), (D.46), and (D.51) together, we conclude that

$$\sup_{x \in \left[\frac{1}{b'}, \frac{1}{a'}\right]} \big|\Psi_n(x) - \Psi_{c_n,H_n}(x)\big| \leq \frac{2}{a'} \sup_{x \in [a',b']} \big|F_n(x) - F_{c_n,H_n}(x)\big|$$

$$\mathbb{E} \sup_{x \in \left[\frac{1}{b'}, \frac{1}{a'}\right]} \big|\Psi_n(x) - \Psi_{c_n,H_n}(x)\big| \leq \frac{2}{a'}\, \mathbb{E} \sup_{x \in [a',b']} \big|F_n(x) - Fc_n, H_n(x)\big| = O\left(\frac{1}{n^{2/5}}\right) . \blacksquare$$

Lemma D.7. *Under Assumptions 3.1-3.2, when $x_n \in (0, +\infty)$ converges to some $x \in [1/b', 1/a']$, we have*

$$\psi_{c_n,H_n}(x_n) - \psi_{c,H}(x_n) \longrightarrow 0 . \tag{D.52}$$

Proof of Lemma D.7. Lemma D.7 is a direct consequence of Jing et al.'s (2010) Lemma 2 once we notice that

$$\psi_{c_n,H_n}(x_n) = -\frac{1}{x_n} f_{c_n,H_n}(1/x_n) \quad \text{and} \quad \psi_{c,H}(x_n) = -\frac{1}{x_n} f_{c,H}(1/x_n) . \blacksquare \tag{D.53}$$

Lemma D.8. *If g is a density with bounded support whose Hilbert transform $\mathcal{H}_g$ exists and is continuous, then*

$$\lim_{x \to \pm\infty} x\mathcal{H}_g(x) = -\frac{1}{\pi} .$$

Proof of Lemma D.8. Let $c_1 := \min \mathsf{Supp}(g)$ and $c_2 := \max \mathsf{Supp}(g)$. Take $d_1 < c_1$ and $d_2 > c_2$.

$$\forall x > d_2 \qquad \pi x\mathcal{H}_g(x) = \int_{c_1}^{c_2} x\frac{g(t)}{t-x}dt \leq \frac{x}{c_1 - x}\int_{c_1}^{c_2} g(t)dt = \frac{x}{c_1 - x}$$

$$\text{and} \quad \pi x\mathcal{H}_g(x) = \int_{c_1}^{c_2} x\frac{g(t)}{t-x}dt \geq \frac{x}{c_2 - x}\int_{c_1}^{c_2} g(t)dt = \frac{x}{c_2 - x} .$$

Since both bounds converge to -1 as $x \to +\infty$, it holds that $\lim_{x\to+\infty} \pi x\mathcal{H}_g(x) = -1$.

$$\forall x < d_1 \qquad \pi x\mathcal{H}_g(x) = \int_{c_1}^{c_2} x\frac{g(t)}{t-x}dt \leq \frac{x}{c_2 - x}\int_{c_1}^{c_2} g(t)dt = \frac{x}{c_2 - x}$$

$$\text{and} \quad \pi x\mathcal{H}_g(x) = \int_{c_1}^{c_2} x\frac{g(t)}{t-x}dt \geq \frac{x}{c_1 - x}\int_{c_1}^{c_2} g(t)dt = \frac{x}{c_1 - x} .$$

Since both bounds converge to -1 as $x \to -\infty$, it holds that $\lim_{x\to-\infty} \pi x\mathcal{H}_g(x) = -1$. $\blacksquare$

Corollary D.1. *If g is a density with bounded support whose Hilbert transform $\mathcal{H}_g$ exists and is continuous, then there exists $C > 0$ such that $\forall x \in \mathbb{R}$, $\big|x\mathcal{H}_g(x)\big| \leq C$.*

D.4 Consistency Result

We next state and prove a theorem about the uniform consistency of the kernel estimators of the first incomplete moment of the spectral precision e.d.f. ψ and its Hilbert transform.

Theorem D.1. *Under Assumptions 3.1-3.2, D.1, and D.2, both*

$$\sup_{x\in[1/b',1/a']}\left|\widehat{\psi}_n(x)-\psi(x)\right|\longrightarrow 0 \quad \textit{and} \quad \sup_{x\in[1/b',1/a']}\left|\mathcal{H}_{\widehat{\psi}_n}(x)-\mathcal{H}_{\psi}(x)\right|\xrightarrow{p} 0\ , \tag{D.54}$$

where the symbol $\xrightarrow{p}$ denotes convergence in probability.

Proof of Theorem D.1. Without loss of generality, in the following developments, we will work on a set of probability one on which F_n converges almost surely to F. We then take n large enough such that both $F_n(a')=0$ and $F_n(b')=1$, which can be done by the results of Bai and Silverstein (1998). First, we claim that

$$\sup_{x\in[a',b']}\left|\breve{m}_{\widehat{\Psi}_n}(x)-\int_{1/b'}^{1/a'}\frac{1}{th_n}\breve{m}_K\left(\frac{x-t}{th_n}\right)d\Psi_{c_n,H_n}(t)\right|\xrightarrow{p}0\ . \tag{D.55}$$

From Equations (D.4)–(D.5) we compute the Stieltjes transform as

$$\breve{m}_{\widehat{\Psi}_n}(x)=\frac{1}{p}\sum_{j=1}^{p}\frac{\lambda_{n,j}^{-1}}{h_n\lambda_{n,j}^{-1}}\breve{m}_K\left(\frac{x-\lambda_{n,j}^{-1}}{h_n\lambda_{n,j}^{-1}}\right)=\int_{1/b'}^{1/a'}\frac{1}{th_n}\breve{m}_K\left(\frac{x-t}{th_n}\right)d\widehat{\Psi}_n(t)\ , \tag{D.56}$$

for all $x\in\mathbb{R}$. Now by using integration by parts, we obtain

$$\begin{aligned}
&\mathbb{E}\sup_{x\in\left[\frac{1}{b'},\frac{1}{a'}\right]}\left|\int_{1/b'}^{1/a'}\frac{1}{th_n}\breve{m}_K\left(\frac{x-t}{th_n}\right)d\Psi_n(t)-\int_{1/b'}^{1/a'}\frac{1}{th_n}\breve{m}_K\left(\frac{x-t}{th_n}\right)d\Psi_{c_n,H_n}(t)\right| \qquad \text{(D.57)}\\
&\quad=\mathbb{E}\sup_{x\in\left[\frac{1}{b'},\frac{1}{a'}\right]}\left|\int_{1/b'}^{1/a'}\frac{1}{th_n}\breve{m}_K\left(\frac{x-t}{th_n}\right)\left[d\Psi_n(t)-d\Psi_{c_n,H_n}(t)\right]\right|\\
&\quad=\mathbb{E}\sup_{x\in\left[\frac{1}{b'},\frac{1}{a'}\right]}\left|\int_{1/b'}^{1/a'}\frac{1}{t^2h_n}\left[\breve{m}_K\left(\frac{x-t}{th_n}\right)+\frac{x}{th_n}\breve{m}'_K\left(\frac{x-t}{th_n}\right)\right]\times\left[\Psi_n(t)-\Psi_{c_n,H_n}(t)\right]dt\right|\\
&\quad=\mathbb{E}\sup_{x\in\left[\frac{1}{b'},\frac{1}{a'}\right]}\left|\int_{\frac{a'x-1}{h_n}}^{\frac{b'x-1}{h_n}}\frac{(1+uh_n)^2}{x^2h_n}\left[\breve{m}_K(u)+\frac{1+uh_n}{h_n}\breve{m}'_K(u)\right]\right.\\
&\qquad\qquad\left.\times\left[\Psi_n\left(\frac{x}{1+uh_n}\right)-\Psi_{c_n,H_n}\left(\frac{x}{1+uh_n}\right)\right]\frac{xh_n}{(1+uh_n)^2}du\right|\\
&\quad=\mathbb{E}\sup_{x\in\left[\frac{1}{b'},\frac{1}{a'}\right]}\left|\int_{\frac{a'x-1}{h_n}}^{\frac{b'x-1}{h_n}}\frac{1}{x}\left[\breve{m}_K(u)+\frac{1+uh_n}{h_n}\breve{m}'_K(u)\right]\right.\\
&\qquad\qquad\left.\times\left[\Psi_n\left(\frac{x}{1+uh_n}\right)-\Psi_{c_n,H_n}\left(\frac{x}{1+uh_n}\right)\right]du\right|\\
&\quad\le\frac{1}{h_n}\mathbb{E}\sup_x\left|\Phi_n(x)-\Psi_{c_n,H_n}(x)\right|\times b'\left[h_n\int_{\frac{a'-b'}{a'h_n}}^{\frac{b'-a'}{a'h_n}}\left|\breve{m}_K(u)\right|du+\frac{b'}{a'}\int_{-\infty}^{+\infty}\left|\breve{m}'_K(u)\right|du\right]\\
&\quad=O\left(\frac{1}{n^{2/5}h_n}\right)\longrightarrow 0\ , \qquad \text{(D.58)}
\end{aligned}$$

where we have used Lemma D.6 in the last line, together with the fact that the multiplier between the square brackets is $O(1)$. This is because this multiplier is the sum of two terms that are each $O(1)$: the first term by Lemma D.1, and the second one by Equation (D.13).

The next aim is to show that

$$\int_{1/b'}^{1/a'} \frac{1}{th_n} \breve{m}_K\left(\frac{x-t}{th_n}\right) d\Psi_{c_n,H_n}(t) - \int_{1/b'}^{1/a'} \frac{1}{th_n} \breve{m}_K\left(\frac{x-t}{th_n}\right) d\Psi_{c,H}(t) \longrightarrow 0 \tag{D.59}$$

uniformly in $x \in [a', b']$. Using the change of variable $u = (x-t)/(th_n)$, this is equivalent to proving that, for any sequence $\{x_n, n \geq 1\}$ in $[a', b']$ converging to x,

$$\int_{(a'x_n-1)/h_n}^{(b'x_n-1)/h_n} \frac{1}{1+uh_n} \breve{m}_K(u) \left[\Psi'_{c_n,H_n}\left(\frac{x_n}{1+uh_n}\right) - \Psi'_{c,H}\left(\frac{x_n}{1+uh_n}\right)\right] du \longrightarrow 0 \ . \tag{D.60}$$

Notice that the first term under the integral, $(1+uh_n)^{-1}$, is bounded from below by $1/b'^2 > 0$ and from above by $1/a'^2 < +\infty$. From Lemma D.4, $\Psi'_{c,H}$ is uniformly bounded on $[a', b']$. Therefore, (D.60) follows from the dominated convergence theorem, Lemma D.5, and Lemma D.7.

The final step is divided into two sub-items, by considering the real part of the Stieltjes transform (which is π times the Hilbert transform of the density) and its imaginary part (which is π times the density itself) separately. Recall that PV denotes the Cauchy principal value of an improper integral. Regarding the real part, we observe that

$$\begin{aligned}
\int_{1/b'}^{1/a'} \frac{1}{th_n} \mathsf{Re}\left[\breve{m}_K\left(\frac{x-t}{th_n}\right)\right] d\Psi_{c,H}(t) &= \int_{1/b'}^{1/a'} \frac{1}{th_n}\, PV \int \frac{k(u)}{u - \frac{x-t}{th_n}} du\, d\Psi_{c,H}(t) \\
&= \int k(u)\, PV \int_{1/b'}^{1/a'} \frac{1}{th_n} \frac{\Psi'_{c,H}(t)}{u - \frac{x-t}{th_n}} dt du \\
&= PV \int \frac{1}{1+uh_n} k(u)\, PV \int_{1/b'}^{1/a'} \frac{\Psi'_{c,H}(t)}{t - \frac{x}{1+uh_n}} dt du \\
&= PV \int \frac{1}{1+uh_n} k(u) \mathsf{Re}\left[\breve{m}_{\Psi_{c,H}}\left(\frac{x}{1+uh_n}\right)\right] du \ .
\end{aligned} \tag{D.61}$$

Therefore,

$$\begin{aligned}
&\sup_{x \in \left[\frac{1}{b'}, \frac{1}{a'}\right]} \left| \int_{1/b'}^{1/a'} \frac{1}{th_n} \mathsf{Re}\left[\breve{m}_K\left(\frac{x-t}{th_n}\right)\right] d\Psi_{c,H}(t) - \mathsf{Re}[\breve{m}_{\Psi_{c,H}}(x)] \right| \\
&\quad = \sup_{x \in \left[\frac{1}{b'}, \frac{1}{a'}\right]} \left| PV \int \frac{1}{1+uh_n} k(u) \mathsf{Re}\left[\breve{m}_{\Psi_{c,H}}\left(\frac{x}{1+uh_n}\right)\right] du - \mathsf{Re}[\breve{m}_{\Psi_{c,H}}(x)] \right| \\
&\quad \leq \sup_{x \in \left[\frac{1}{b'}, \frac{1}{a'}\right]} \left| PV \int \frac{1}{1+uh_n} k(u) \left\{ \mathsf{Re}\left[\breve{m}_{\Psi_{c,H}}\left(\frac{x}{1+uh_n}\right)\right] - \mathsf{Re}[\breve{m}_{\Psi_{c,H}}(x)] \right\} du \right| \\
&\qquad + \sup_{x \in \left[\frac{1}{b'}, \frac{1}{a'}\right]} \left| \mathsf{Re}[\breve{m}_{\Psi_{c,H}}(x)] \right| \times \left| 1 - PV \int \frac{1}{1+uh_n} k(u) du \right| \ .
\end{aligned} \tag{D.62}$$

Note that, by Lemma D.3,

$$\lim_{n \to \infty} PV \int \frac{1}{1+uh_n} k(u) du = \int k(u) du = 1 \ . \tag{D.63}$$

Therefore, using also Remark D.2, the second term in (D.62) is $o(1)$. The first term is more complicated, and needs to be split into three parts, for which the arguments are different.

First, using Remark D.2 again, there exists some finite constant M that is an upper bound on $|\mathsf{Re}[\breve{m}_{\Psi_{c,H}}(x)|$ for all $x \in \mathbb{R}$. For any small $\varepsilon > 0$, there exists some sufficiently large $R > 0$ such that

$$\forall n \in \mathbb{N} \qquad \int_R^{+\infty} \frac{1}{1+uh_n} k(u)du \leq \int_R^{+\infty} k(u)du = 1 - K(R) \leq \frac{\varepsilon}{2M} \ . \tag{D.64}$$

Then:

$$\begin{aligned} \limsup_{n\to\infty} \sup_{x\in\left[\frac{1}{b'},\frac{1}{a'}\right]} \left| \int_R^{+\infty} \frac{1}{1+uh_n} k(u) \left\{ \mathsf{Re}\left[\breve{m}_{\Psi_{c,H}}\left(\frac{x}{1+uh_n} \right) \right] - \mathsf{Re}[\breve{m}_{\Psi_{c,H}}(x)] \right\} du \right| \\ \leq 2M \limsup_{n\to\infty} \int_R^{+\infty} \frac{1}{1+uh_n} k(u)du \leq \varepsilon \ . \end{aligned} \tag{D.65}$$

Second,

$$\begin{aligned} &\sup_{x\in\left[\frac{1}{b'},\frac{1}{a'}\right]} \left| \int_{-R}^{R} \frac{1}{1+uh_n} k(u) \left\{ \mathsf{Re}\left[\breve{m}_{\Psi_{c,H}}\left(\frac{x}{1+uh_n} \right) \right] - \mathsf{Re}[\breve{m}_{\Psi_{c,H}}(x)] \right\} du \right| \\ &\quad \leq \sup_{x\in\left[\frac{1}{b'},\frac{1}{a'}\right]} \int_{-R}^{R} \frac{1}{1+uh_n} k(u) \left| \mathsf{Re}\left[\breve{m}_{\Psi_{c,H}}\left(\frac{x}{1+uh_n} \right) \right] - \mathsf{Re}[\breve{m}_{\Psi_{c,H}}(x)] \right| du \\ &\quad \leq \sup_{\substack{x,y\in\left[\frac{1}{b'}-\frac{Rh_n}{1-Rh_n},\frac{1}{a'}+\frac{Rh_n}{1-Rh_n}\right] \\ |x-y|\leq\frac{Rh_n}{1-Rh_n}}} \left| \mathsf{Re}\left[\breve{m}_{\Psi_{c,H}}(y) \right] - \mathsf{Re}[\breve{m}_{\Psi_{c,H}}(x)] \right| \times \left| \int_{-R}^{R} \frac{1}{1+uh_n} k(u)du \right| \\ &\quad \leq \sup_{\substack{x,y\in\left[\frac{1}{2b'},\frac{2}{a'}\right] \\ |x-y|\leq\frac{Rh_n}{1-Rh_n}}} \left| \mathsf{Re}\left[\breve{m}_{\Psi_{c,H}}(y) \right] - \mathsf{Re}[\breve{m}_{\Psi_{c,H}}(x)] \right| \times \left| \int_{-R}^{R} \frac{1}{1+uh_n} k(u)du \right| \end{aligned} \tag{D.66}$$

for sufficiently large n. The first term of expression (D.66) converges to zero thanks to the Heine-Cantor theorem, and the second term remains bounded because, when n is large enough for h_n to be below $\frac{1}{2R}$,

$$\left| \int_{-R}^{R} \frac{1}{1+uh_n} k(u)du \right| = \int_{-R}^{R} \frac{1}{1+uh_n} k(u)du \leq \int_{-R}^{R} \frac{1}{2} k(u)du \leq \frac{1}{2} \ . \tag{D.67}$$

This guarantees that the upper bound (D.66) is $o(1)$ as $n \to \infty$.

Third,

$$\begin{aligned} &\limsup_{n\to\infty} \sup_{x\in\left[\frac{1}{b'},\frac{1}{a'}\right]} \left| PV \int_{-\infty}^{-R} \frac{1}{1+uh_n} k(u) \left\{ \mathsf{Re}\left[\breve{m}_{\Psi_{c,H}}\left(\frac{x}{1+uh_n} \right) \right] - \mathsf{Re}[\breve{m}_{\Psi_{c,H}}(x)] \right\} du \right| \\ &\qquad \leq \limsup_{n\to\infty} \sup_{x\in\left[\frac{1}{b'},\frac{1}{a'}\right]} \left| PV \int_{-\infty}^{-R} \frac{1}{1+uh_n} k(u) \mathsf{Re}\left[\breve{m}_{\Psi_{c,H}}\left(\frac{x}{1+uh_n} \right) \right] du \right| \\ &\qquad\quad + \limsup_{n\to\infty} \sup_{x\in\left[\frac{1}{b'},\frac{1}{a'}\right]} \left| PV \int_{-\infty}^{-R} \frac{1}{1+uh_n} k(u) \mathsf{Re}[\breve{m}_{\Psi_{c,H}}(x)] du \right| \\ &\qquad =: \limsup_{n\to\infty} A_n + \limsup_{n\to\infty} B_n \ . \end{aligned}$$

Concerning the first term A_n, for any sufficiently large n and any $x \in \left[\frac{1}{b'}, \frac{1}{a'}\right]$, define the function

$$g_{n,x}(u) := \begin{cases} \dfrac{1}{1+uh_n} \mathsf{Re}\left[\breve{m}_{\Psi_{c,H}}\left(\dfrac{x}{1+uh_n}\right)\right] & \text{if } u \neq -1/h_n \\ -\dfrac{1}{x} & \text{if } u = -1/h_n \ . \end{cases}$$

Consider the change of variables $v := x/(1+uh_n)$. Then as $u \downarrow -1/h_n$, $v \to +\infty$. From the Lemma D.8, we get $\lim_{v\to+\infty} v\mathsf{Re}\left[\breve{m}_{\Psi_{c,H}}(v)\right] = -1$, which implies that $\lim_{u\downarrow -1/h_n} x g_{n,x}(u) = -1$. Similarly, $\lim_{u\uparrow -1/h_n} x g_{n,x}(x) = \lim_{v\to-\infty} v\mathsf{Re}\left[\breve{m}_{\Psi_{c,H}}(v)\right] = -1$, hence the function $g_{n,x}$ is continuous over $u \in \mathbb{R}$. As a result, we can get rid of the Cauchy principal value:

$$PV \int_{-\infty}^{-R} \frac{1}{1+uh_n} k(u) \mathsf{Re}\left[\breve{m}_{\Psi_{c,H}}\left(\frac{x}{1+uh_n}\right)\right] du = \int_{-\infty}^{-R} g_{n,x}(u)k(u)du \ . \tag{D.68}$$

Note that for n sufficiently large and for all $x \in \left[\frac{1}{b'}, \frac{1}{a'}\right]$

$$g_{n,x}(u) = \begin{cases} \pi v \mathcal{H}_{\psi_{c,H}}(v)/x & \text{if } u \neq -1/h_n \\ -1/x & \text{if } u = -1/h_n \ . \end{cases}$$

By Corollary D.1, there exists a finite constant $K > 0$ such that for all sufficiently large n, for all $x \in \left[\frac{1}{a'}, \frac{1}{b'}\right]$ and for all $u \in \mathbb{R}$, $\left|g_{n,x}(u)\right| \leq \pi K/a' =: \Delta$. This result, combined with (D.68), implies that for sufficiently large n

$$\sup_{x\in\left[\frac{1}{a'},\frac{1}{b'}\right]} \left| PV \int_{-\infty}^{-R} \frac{1}{1+uh_n} k(u) \mathsf{Re}\left[\breve{m}_{\Psi_{c,H}}\left(\frac{x}{1+uh_n}\right)\right] du \right| \leq \Delta \int_{-\infty}^{-R} k(u)du \ . \tag{D.69}$$

Without loss of generality we can consider that R has been chosen sufficiently large to ensure that this upper bound is less than ε.

Concerning the second term B_n, it holds that

$$\begin{aligned} B_n &\leq M \left| PV \int_{-\infty}^{-R} \frac{1}{1+uh_n} k(u)du \right| \\ &\leq M \left| PV \int_{-\infty}^{+\infty} \frac{1}{1+uh_n} k(u)du - \int_{-R}^{+\infty} \frac{1}{1+uh_n} k(u)du \right| \end{aligned}$$

By Lemma D.3,

$$\lim_{n\to\infty} PV \int_{-\infty}^{+\infty} \frac{1}{1+uh_n} k(u)du = 1 \ .$$

By the dominated convergence theorem, there exists some $R' > 0$ sufficiently large such that

$$1 \geq \lim_{n\to\infty} \int_{-R'}^{+\infty} \frac{1}{1+uh_n} k(u)du \geq 1 - \frac{\varepsilon}{C_2} \ .$$

Without loss of generality it can be assumed that $R' = R$. Therefore,

$$\limsup_{n\to\infty} B_n \leq M \left|1 - \left(1 - \frac{\varepsilon}{M}\right)\right| = \varepsilon \ . \tag{D.70}$$

Concerning the proof for the imaginary part, the statement we seek to establish is

$$\sup_{x \in \left[\frac{1}{a'}, \frac{1}{b'}\right]} \left| \int \frac{1}{t h_n} \mathsf{Im}\left[\breve{m}_K\left(\frac{x-t}{h_n}\right)\right] d\Psi_{c,H}(t) - \mathsf{Im}[\breve{m}_{\Psi_{c,H}}(x)] \right| \longrightarrow 0 \ . \tag{D.71}$$

A closely related statement, namely

$$\sup_{x \in [a,b]} \left| \int \frac{1}{h_n} \mathsf{Im}\left[\breve{m}_K\left(\frac{x-t}{h_n}\right)\right] dF_{c,H}(t) - \mathsf{Im}[\breve{m}_{F_{c,H}}(x)] \right| \longrightarrow 0 \ , \tag{D.72}$$

was proven by Jing et al. (2010) in the course of proving their Theorem 1 at the end of Section 5.1. It can be verified that the methods developed in this paper to simultaneously handle (i) locally adaptive bandwidth, (ii) the first incomplete moment function of the spectral e.d.f. of the precision matrix, and (iii) unbounded kernel support, can be adapted to establish the truth of (D.71), using the techniques developed above for the real part. Indeed, the real part was a much tougher nut to crack because it also required moving from the spectral density to its Hilbert transform; whereas the imaginary part is just the spectral density itself (up to rescaling by π).

Note that (D.66) and (D.71) together imply

$$\sup_{x \in \left[\frac{1}{a'}, \frac{1}{b'}\right]} \left| \int \frac{1}{t h_n} \breve{m}_K\left(\frac{x-t}{h_n}\right) d\Psi_{c,H}(t) - \breve{m}_{\Psi_{c,H}}(x) \right| \longrightarrow 0 \ . \tag{D.73}$$

Results (D.55), (D.59), and (D.73) together conclude the proof of Theorem D.1. ■

D.5 Proof of Theorem 3.1

Theorem 3.1 of Ledoit and Wolf (2018) shows that, under Assumptions 3.1–3.3, the quantity $\mathcal{L}_n^{\mathrm{ST}}(\Sigma_n, \widetilde{\Sigma}_n)$, which represents Stein's loss for any covariance matrix estimator $\widetilde{\Sigma}_n$ in the rotation-equivariant class of Definition 3.2, converges almost surely as p and n go to infinity together to the nonrandom limit:

$$\sum_{k=1}^{\kappa} \int_{a_k}^{b_k} \left\{ \frac{1 - c - 2\,c\,x\,\mathsf{Re}[\breve{m}_F(x)]}{x} \widetilde{\delta}(x) - \log\left[\widetilde{\delta}(x)\right] \right\} dF(x) + \int_{-\infty}^{+\infty} \log(t)\, dH(t) - 1 \ . \tag{D.74}$$

From Corollary 3.1.a of Ledoit and Wolf (2018), it then follows that a covariance matrix estimator $\widetilde{S}_n$ minimizes in the class of rotation-equivariant estimators described in Definition 3.2 the almost sure limit (D.74) of Stein's loss if and only if its limiting shrinkage function $\widetilde{\delta}$ verifies $\forall x \in \mathsf{Supp}(F)$, $\widetilde{\delta}(x) = \delta^{\mathrm{ST}}(x)$, where

$$\forall x \in \mathsf{Supp}(F) \qquad \delta^{\mathrm{ST}}(x) \coloneqq \frac{x}{1 - c - 2\,c\,x\,\mathsf{Re}[\breve{m}_F(x)]} \ . \tag{D.75}$$

By Equation (D.35), we know that

$$\forall x \in \mathsf{Supp}(F) \qquad \mathsf{Re}[\breve{m}_\Psi(1/x)] = -x\,\mathsf{Re}[\breve{m}_F(x)] \ . \tag{D.76}$$

Substituting Equation (D.76) into Equation (D.75) yields

$$\forall x \in \mathsf{Supp}(F) \qquad \delta^{\mathrm{ST}}(x) = \frac{x}{1 - c + 2\,c\,\mathsf{Re}[\breve{m}_{\Psi}(1/x)]} \tag{D.77}$$

$$\delta^{\mathrm{ST}}(x) = \frac{1}{(1-c)x^{-1} + cx^{-1} \times 2\pi\mathcal{H}_{\psi}(1/x)} \; . \tag{D.78}$$

Consider the shrinkage function

$$\forall x \in (0, +\infty) \qquad \widehat{d}_n(x) := \frac{1}{\left(1 - \frac{p}{n}\right) x^{-1} + \left(\frac{p}{n}\right) x^{-1} \times 2\widehat{\theta}_n\left(x^{-1}\right)} \tag{D.79}$$

A comparison with the estimator proposed in Theorem 3.1 reveals that

$$\forall i = 1, \ldots, p \qquad \widehat{d}_{n,i} = \widehat{d}_n\left(\lambda_{n,i}\right) \; . \tag{D.80}$$

Thanks to Equation (D.12), the shrinkage function can be rewritten as

$$\forall x \in (0, +\infty) \qquad \widehat{d}_n(x) = \frac{1}{\left(1 - \frac{p}{n}\right) x^{-1} + \left(\frac{p}{n}\right) x^{-1} \times 2\pi\mathcal{H}_{\widehat{\Psi}_n}\left(x^{-1}\right)} \; . \tag{D.81}$$

This is where Theorem D.1 comes into play: it implies that

$$\forall x \in \mathsf{Supp}(F) \qquad \widehat{d}_n(x) \xrightarrow{p} \delta^{\mathrm{ST}}(x) \; . \tag{D.82}$$

Combined with the aforementioned Corollary 3.1.a of Ledoit and Wolf (2018), this concludes the proof of Theorem 3.1. ■

D.6 Proof of Theorem 4.1

Theorem 4.2 of Ledoit and Wolf (2018) shows that, under Assumptions 3.1–3.3, the quantity $\mathcal{L}_n^{\mathrm{FR}}(\Sigma_n, \widetilde{\Sigma}_n)$, which represents the Frobenius loss for any covariance matrix estimator $\widetilde{\Sigma}_n$ in the rotation-equivariant class of Definition 3.2, converges almost surely as p and n go to infinity together to the nonrandom limit:

$$\int_{-\infty}^{+\infty} x^2\, dH(x) + \sum_{k=1}^{\kappa} \left\{ -2 \int_{a_k}^{b_k} \frac{x\,\widetilde{\delta}(x)}{\left|1 - c - c\,x\,\breve{m}_F(x)\right|^2} dF(x) + \int_{a_k}^{b_k} \widetilde{\delta}(x)^2 dF(x) \right\} \; . \tag{D.83}$$

From Corollary 4.2 of Ledoit and Wolf (2018), it then follows that a covariance matrix estimator $\widetilde{S}_n$ minimizes in the class of rotation-equivariant estimators described in Definition 3.2 the almost sure limit (D.83) of the Frobenius loss if and only if its limiting shrinkage function $\widetilde{\delta}$ verifies

$\forall x \in \mathsf{Supp}(F)$, $\widetilde{\delta}(x) = \delta^{\mathrm{FR}}(x)$, where $\forall x \in \mathsf{Supp}(F)$

$$\delta^{\mathrm{FR}}(x) \coloneqq \frac{x}{\left|1 - c - c\,x\,\breve{m}_F(x)\right|^2} \tag{D.84}$$

$$= \frac{x}{\left\{1 - c - c\,x\,\mathsf{Re}\left[\breve{m}_F(x)\right]\right\}^2 + c^2\,x^2\,\mathsf{Im}\left[\breve{m}_F(x)\right]^2} \tag{D.85}$$

$$= \frac{x}{(1-c)^2 - 2c(1-c)x\,\mathsf{Re}\left[\breve{m}_F(x)\right] + c^2\left\{x^2\,\mathsf{Re}\left[\breve{m}_F(x)\right]^2 + x^2\,\mathsf{Im}\left[\breve{m}_F(x)\right]^2\right\}}$$

$$= \frac{x}{(1-c)^2 + 2c(1-c)\mathsf{Re}\left[\breve{m}_\Psi(x^{-1})\right] + c^2\left\{\mathsf{Re}\left[\breve{m}_\Psi(x^{-1})\right]^2 + \mathsf{Im}\left[\breve{m}_\Psi(x^{-1})\right]^2\right\}}$$

$$= \frac{1}{(1-c)^2\,x^{-1} + 2c(1-c)x^{-1}\pi\mathcal{H}_\psi\left(x^{-1}\right) + c^2x^{-1}\pi^2\left[\mathcal{H}_\psi\left(x^{-1}\right)^2 + \psi\left(x^{-1}\right)^2\right]}$$

Consider the shrinkage function defined for all $x \in (0, +\infty)$ by

$$\widehat{\delta}_n(x) \coloneqq \frac{1}{\left(1 - \frac{p}{n}\right)^2 x^{-1} + 2\frac{p}{n}\left(1 - \frac{p}{n}\right) x^{-1}\widehat{\theta}_n\left(x^{-1}\right) + \left(\frac{p}{n}\right)^2 x^{-1}\left[\widehat{\theta}_n\left(x^{-1}\right)^2 + \mathcal{H}_{\widehat{\theta}_n}\left(x^{-1}\right)^2\right]}$$

A comparison with the estimator proposed in Theorem 4.1 reveals that

$$\forall i = 1, \ldots, p \qquad \widehat{\delta}_{n,i} = \widehat{\delta}_n\left(\lambda_{n,i}\right) \ . \tag{D.86}$$

By Equation (D.12) and the anti-involution property of the Hilbert transform, the functions $\pi\widehat{\psi}_n(x)$ and $\widehat{\theta}_n(x)$ constitute a Hilbert pair. This means that $\mathcal{H}_{\widehat{\psi}_n}(x) = \widehat{\theta}_n(x)/\pi$ and $\mathcal{H}_{\widehat{\theta}_n}(x) = -\pi\widehat{\psi}_n(x)$. With this in mind, the shrinkage function $\widehat{\delta}_n(x)$ can be rewritten as

$$\frac{1}{\left(1 - \frac{p}{n}\right)^2 x^{-1} + 2\frac{p}{n}\left(1 - \frac{p}{n}\right) x^{-1}\pi\mathcal{H}_{\widehat{\psi}_n}\left(x^{-1}\right) + \left(\frac{p}{n}\right)^2 x^{-1}\pi^2\left[\mathcal{H}_{\widehat{\psi}_n}\left(x^{-1}\right)^2 + \widehat{\psi}_n\left(x^{-1}\right)^2\right]}$$

Then Theorem D.1 implies that

$$\forall x \in \mathsf{Supp}(F) \qquad \widehat{\delta}_n(x) \xrightarrow{p} \delta^{\mathrm{FR}}(x) \ . \tag{D.87}$$

Combined with the aforementioned Corollary 4.2 of Ledoit and Wolf (2018), this concludes the proof of Theorem 4.1. ■

E Singular Case

The sample covariance matrix is singular when the dimension p exceeds the sample size n. Its $p - n$ smallest eigenvalues $(\lambda_{n,1}, \ldots, \lambda_{n,p-n})$ are then all equal to zero. In this case, the finite-sample optimal estimator for $\mathcal{L}^{\mathrm{MV}}$, $\mathcal{L}^{\mathrm{FR}}$ and $\mathcal{L}^{\mathrm{IS}}$ treats the null space differently:

$$\forall i = 1, \ldots, p - n \qquad \overline{d}_{n,i} \coloneqq \frac{1}{p-n}\sum_{j=1}^{p-n} u_{n,j}'\Sigma_n u_{n,j} \tag{E.1}$$

$$\forall i = p - n + 1, \ldots, p \qquad \overline{d}_{n,i} \coloneqq u_{n,i}'\Sigma_n u_{n,i} \tag{E.2}$$

The counterpart to Assumption 3.1 is:

Assumption E.1 (Singular Case)**.** *Let n denote the sample size and $p := p(n)$ the number of variables. It is assumed that the concentration (ratio) $c_n := p/n$ converges, as $n \to \infty$, to a limiting concentration ratio $c \in (1, +\infty)$. Furthermore, there exists a compact interval included in $(1, +\infty)$ that contains p/n for all n large enough.*

E.1 Working With Non-Null Sample Eigenvalues

When $p > n$, or asymptotically $c \in (1, +\infty)$, it is more mathematically judicious to work with the e.d.f. of the non-null sample eigenvalues

$$\forall x \in \mathbb{R} \qquad \underline{F}_n(x) := \frac{1}{n} \sum_{i=p-n+1}^{p} \mathbf{1}_{\{\lambda_{n,i} \leq x\}} = \frac{p}{n} F_n(x) - \frac{p-n}{n} \mathbf{1}_{\{x \geq 0\}} \ . \tag{E.3}$$

Under Assumptions E.1, 3.2 and 3.3, Theorem 1.1 of Silverstein (1995) shows that

$$\forall x \in \mathbb{R} \quad \underline{F}_n(x) \xrightarrow{\text{a.s.}} \underline{F}(x) \quad \text{where} \quad \underline{F}(x) := cF(x) - (c-1)\mathbf{1}_{\{x \geq 0\}} \ . \tag{E.4}$$

Theorem 1.1 of Silverstein and Choi (1995) implies that, under the same assumptions, $\underline{F}$ admits a continuous derivative $\underline{f}$ that itself admits a well-defined and continuous Hilbert transform $\mathcal{H}_{\underline{f}}$; or equivalently, that $\breve{m}_{\underline{F}}$ exists and is continuous. The relationship between the two Stieltjes transforms is

$$\forall x \in \mathbb{R} \qquad \breve{m}_{\underline{F}}(x) = c\breve{m}_F(x) + \frac{c-1}{x} \ . \tag{E.5}$$

The relationship between the corresponding densities and their Hilbert transforms follows naturally by taking the real and imaginary parts as per Equation (C.7).

We can do the same with the inverses of the non-null eigenvalues. If the sample precision matrix existed when $p > n$, these would be its finite eigenvalues.

$$\forall x \in \mathbb{R} \qquad \underline{\Phi}_n(x) := \frac{1}{n} \sum_{i=p-n+1}^{p} \mathbf{1}_{\{\lambda_{n,i}^{-1} \leq x\}} = \begin{cases} 1 - \underline{F}_n(1/x) & \text{if } x > 0 \\ 0 & \text{otherwise .} \end{cases} \tag{E.6}$$

Under Assumptions E.1 and 3.2, for all $x \in \mathbb{R}$, $\underline{\Phi}_n(x)$ converges almost surely to

$$\underline{\Phi}(x) := \begin{cases} 1 - \underline{F}(1/x) & \text{if } x > 0 \\ 0 & \text{otherwise .} \end{cases} \tag{E.7}$$

Under the same assumptions, $\underline{\Phi}$ admits a continuous derivative $\underline{\phi}$ that itself admits a well-defined and continuous Hilbert transform $\mathcal{H}_{\underline{\phi}}$; or equivalently, $\breve{m}_{\underline{\Phi}}$ exists and is continuous. Finally, we will also need the first incomplete moment function

$$\forall x \in \mathbb{R} \qquad \underline{\Psi}_n(x) := L\underline{\Phi}_n(x) = \frac{1}{n} \sum_{i=p-n+1}^{p} \lambda_{n,i}^{-1} \mathbf{1}_{\{\lambda_{n,i}^{-1} \leq x\}} \ . \tag{E.8}$$

Under Assumptions E.1 and 3.2, for all $x \in \mathbb{R}$, $\underline{\Psi}_n(x)$ converges almost surely to a limit

$$\underline{\Psi}(x) := L\underline{\Phi}(x) \ , \tag{E.9}$$

which admits a continuous derivative $\underline{\psi}$ whose Hilbert transform $\mathcal{H}_{\underline{\psi}}$ exists and is continuous.

E.2 Preliminary Results

The following results have never been, as such, derived for the Frobenius loss in the singular case, although comparable results have been derived with respect to Stein's loss and the Minimum Variance loss in the singular case, and with respect to the Frobenius loss in the $p < n$ case. So we must build these foundations before proceeding any further.

Definition E.1. *Define* $\forall x \in \mathbb{R},\ \Gamma_n(x) \coloneqq p^{-1} \sum_{i=1}^{p} u_{n,i}' \Sigma_n u_{n,i} \times \mathbf{1}_{[\lambda_{n,i}, +\infty)}(x)$.

Lemma E.1. *Under Assumptions E.1 and 3.2, there exists a nonrandom function* Γ *defined on* $\mathbb{R}$ *such that* $\Gamma_n(x)$ *converges almost surely to* $\Gamma(x)$*, for all* $x \in \mathbb{R} \setminus \{0\}$*. Furthermore,* Γ *is continuously differentiable on* $\mathbb{R} \setminus \{0\}$ *and can be expressed as* $\forall x \in \mathbb{R},\ \Gamma(x) = \int_{-\infty}^{x} \gamma(\lambda) dF(\lambda)$*, where*

$$\forall x \in \mathbb{R} \qquad \gamma(x) \coloneqq \begin{cases} 0 & \text{if } x < 0, \\ \dfrac{1}{(c-1)\breve{m}_{\underline{F}}(0)} & \text{if } x = 0, \\ \dfrac{x}{|1 - c - cx\breve{m}_F(x)|} & \text{if } x > 0\ . \end{cases}$$

Proof of Lemma E.1. The proof of Lemma E.1 follows directly from Ledoit and Péché's (2011) Theorem 4 and the corresponding proof, bearing in mind that we are in the case $c > 1$ (which they call $\gamma < 1$) because of Assumption E.1. ■

Theorem E.1. *Under Assumptions E.1, 3.2, and 3.3,*

$$\mathcal{L}_n^{FR}(\Sigma_n, \widetilde{\Sigma}_n) \xrightarrow{\text{a.s.}} \int_{-\infty}^{+\infty} x^2 dH(x) + \sum_{k=1}^{\kappa} \int_{a_k}^{b_k} \left[\widetilde{\delta}(x)^2 - 2 \frac{\widetilde{\delta}(x)}{x |\breve{m}_{\underline{F}}(x)|^2} \right] dF(x) + \frac{c-1}{c} \left[\widetilde{\delta}(0)^2 - 2 \frac{\widetilde{\delta}(0)}{(c-1)\breve{m}_{\underline{F}}(0)} \right] . \tag{E.10}$$

Proof of Theorem E.1. The proof of Theorem E.1 is similar to the proof of Ledoit and Wolf's (2018) Theorems 4.2 and 6.1, so is omitted. ■

E.3 Shrinkage When $p > n$

Theorem E.10 shows that, under Assumptions E.1, 3.2, and 3.3, the quantity $\mathcal{L}_n^{\text{FR}}(\Sigma_n, \widetilde{\Sigma}_n)$, which represents the Frobenius loss for any covariance matrix estimator $\widetilde{\Sigma}_n$ in the rotation-equivariant class of Definition 3.2, converges almost surely as p and n go to infinity together to the nonrandom limit (E.10). Differentiating with respect to $\widetilde{\delta}(x)$ for any given x and solving the resulting first-order condition yields the optimum

$$\forall x \in \mathbb{R} \qquad \delta^{\text{FR}}(x) \coloneqq \begin{cases} \dfrac{1}{(c-1)\breve{m}_{\underline{F}}(0)} & \text{if } x = 0, \\ \dfrac{x}{|1 - c - cx\breve{m}_F(x)|^2} & \text{if } x > 0\ . \end{cases} \tag{E.11}$$

By arguments parallel to those employed in Corollary 6.1 of Ledoit and Wolf (2018), it then follows that a covariance matrix estimator $\widetilde{S}_n$ minimizes in the class of rotation-equivariant

estimators described in Definition 3.2 the almost sure limit (E.10) of the Frobenius loss if and only if its limiting shrinkage function $\widetilde{\delta}$ verifies $\forall x \in \mathsf{Supp}(F)$, $\widetilde{\delta}(x) = \delta^{\mathrm{FR}}(x)$ as defined by (E.11). Regarding the non-null sample eigenvalues:

$$\forall x \in (0, +\infty) \qquad \delta^{\mathrm{FR}}(x) = \frac{x}{\left|1 - c - c\, x\, \breve{m}_F(x)\right|^2} = \frac{1}{x \left|\breve{m}_{\underline{F}}(x)\right|^2} = \frac{x}{\left|\breve{m}_{\underline{\Psi}}\left(x^{-1}\right)\right|^2} \tag{E.12}$$

$$= \frac{1}{x^{-1}\pi^2 \left[\mathcal{H}_{\underline{\psi}}\left(x^{-1}\right)^2 + \underline{\psi}\left(x^{-1}\right)^2\right]} \ . \tag{E.13}$$

This oracle shrinkage function is estimated by its *bona fide* counterpart

$$\widehat{\delta}_n(x) \coloneqq \frac{1}{x^{-1}\pi^2 \left[\mathcal{H}_{\underline{\widehat{\psi}}_n}\left(x^{-1}\right)^2 + \underline{\widehat{\psi}}_n\left(x^{-1}\right)^2\right]} \ . \tag{E.14}$$

A comparison with the estimator proposed in Section 5 reveals that

$$\forall i = p - n + 1, \ldots, p \qquad \widehat{\delta}_n\left(\lambda_{n,i}\right) = \frac{1}{x^{-1}\left[\underline{\widehat{\theta}}_n\left(\lambda_{n,i}^{-1}\right)^2 + \mathcal{H}_{\underline{\widehat{\theta}}_n}\left(\lambda_{n,i}^{-1}\right)^2\right]} = \widehat{\delta}_{n,i} \ . \tag{E.15}$$

Finally, the null space of the sample covariance matrix requires a separate approach. The shrinkage formula for $\widehat{\delta}_{n,i}$ $(i = 1, \ldots, p - n)$ comes from the following proposition:

Proposition E.1. *Under Assumptions E.1 and 3.2,*

$$\left(\frac{p}{n} - 1\right) \times \frac{1}{n} \sum_{j=p-n+1}^{p} \lambda_{n,j}^{-1} \xrightarrow{a.s.} (c-1)\breve{m}_{\underline{F}}(0) \ . \tag{E.16}$$

Proof of Proposition E.1. Let X be a random variable with distribution (function) F. Then X can be written as

$$X = \frac{c-1}{c} Y + \frac{1}{c} Z \ ,$$

where Y is a random variable corresponding to a point mass at zero and Z is random variable corresponding to the continuous part of F, having support in $[a, b]$ for $0 < a < b < +\infty$. Under the maintained assumptions, for all n large enough, the collection of sample eigenvalues $\{\lambda_{n,p-n+1}, \ldots, \lambda_{n,p}\}$ is also contained in $[a, b]$ almost surely, with the empirical distribution of these n values converging weakly to Z almost surely. This fact implies by the continuous mapping theorem that the empirical distribution of the n values $\{\lambda_{n,p-n+1}^{-1}, \ldots, \lambda_{n,p}^{-1}\}$ converges weakly to Z^{-1} almost surely. Further note that that interval $[1/b, 1/a]$ contains the support of Z^{-1} as well as the collection $\{\lambda_{n,p-n+1}^{-1}, \ldots, \lambda_{n,p}^{-1}\}$, for all n large enough, almost surely. By Skorokhod's representation theorem and the dominated convergence theorem it then follows that

$$\frac{1}{n} \sum_{j=p-n+1}^{p} \lambda_{n,j}^{-1} \xrightarrow{a.s.} \mathbb{E}\big(Z^{-1}\big).$$

The proof is complete by noting that $\breve{m}_{\underline{F}}(0) = \mathbb{E}\big(Z^{-1}\big)$ together with the fact that $p/n \to c$. ■

F Supplementary Monte Carlo Simulations

Further comfort with our approach can be gained by broadening the scope of the numerical inverstigations beyond those already reported in Section 7.

F.1 Post-Processing With the Pool Adjacent Violators (PAV) Algorithm

The shrunken eigenvalues in the QIS estimator are guaranteed to be strictly positive, as mentioned in Remark 4.2. However, they are not guaranteed to preserve ordering. This is a matter worth exploring further, given that Stein's isotonization algorithm, as detailed by Lin and Perlman (1985, pp. 426–428), ensures both positivity and order preservation.

In this section, we revisit the convergence graph in Figure 5. The difference is that we now add the QIS+PAV estimator, which post-processes the QIS eigenvalues through the Pool Adjacent Violators (PAV) algorithm of Ayer et al. (1955), a commonly used method for restoring order. Figure 11 displays the results.

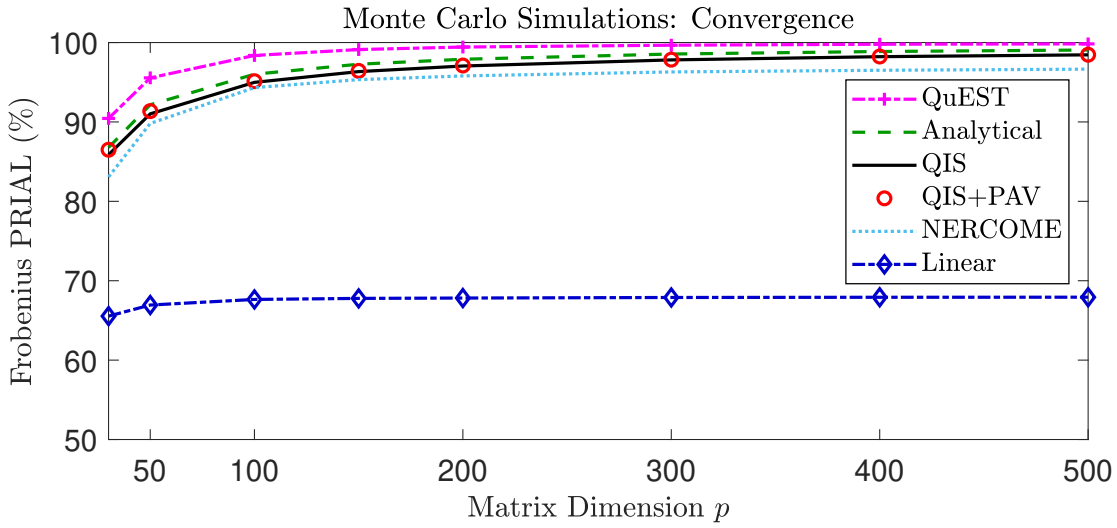

Figure 11: Evolution of the PRIAL as the matrix dimension and the sample size go to infinity together.

One can see that the circles (representing QIS+PAV) are almost exactly superimposed with the solid line (representing QIS); therefore, the two respective performances are essentially the same for practical purposes. In the baseline scenario of Section 7.1, QIS+PAV has a PRIAL of 97.11% vs. 97.04% for simple QIS, a minute difference. Furthermore, since the QIS+PAV estimator belongs to the class in Definition 3.2, Theorem 4.1 applies to it as well, implying that QIS+PAV has the same limiting loss as simple QIS under large-dimensional asymptotics, *not lower*.

Another way to answer the question is to notice that Proposition 4.2 enjoins us to make the ith shrunken eigenvalues very close to $\overline{d}_{n,i} := u_{n,i}' \Sigma_n u_{n,i}$ (for $i = 1, \ldots, p$). These are the eigenvalues of what we call the FSOPT (for finite-sample optimal) estimator in Section 7. Their expectations are in order: $\mathbb{E}[\overline{d}_{n,1}] \leq \mathbb{E}[\overline{d}_{n,2}] \leq \cdots \leq \mathbb{E}[\overline{d}_{n,p}]$, but they themselves need not be. To measure the practical importance of this disorder, we use a procedure analogous to the one

of Ledoit and Wolf (2020, Appendix D.4). We compare the percentage of nearest-neighbor order violations for FSOPT with that for QIS across the eight (p, n) pairs used to generate Figure 11. More formally, these percentages are defined as

$$\frac{1}{p-1}\sum_{i=1}^{p-1}\mathbf{1}_{\{\overline{d}_{n,i}>\overline{d}_{n,i+1}\}} \quad \text{and} \quad \frac{1}{p-1}\sum_{i=1}^{p-1}\mathbf{1}_{\{\widehat{\delta}_{n,i}>\widehat{\delta}_{n,i+1}\}}\,, \tag{F.1}$$

for FSOPT and QIS, respectively. We also record the average magnitudes of the violations, defined as

$$\frac{1}{\overline{a}_n}\sum_{i=1}^{p-1}\frac{\overline{d}_{n,i}-\overline{d}_{n,i+1}}{\frac{1}{2}\left(\overline{d}_{n,i}+\overline{d}_{n,i+1}\right)}\mathbf{1}_{\{\overline{d}_{n,i}>\overline{d}_{n,i+1}\}} \quad \text{and} \quad \frac{1}{\widehat{\alpha}_n}\sum_{i=1}^{p-1}\frac{\widehat{\delta}_{n,i}-\widehat{\delta}_{n,i+1}}{\frac{1}{2}\left(\widehat{\delta}_{n,i}+\widehat{\delta}_{n,i+1}\right)}\mathbf{1}_{\{\widehat{\delta}_{n,i}>\widehat{\delta}_{n,i+1}\}}\,, \tag{F.2}$$

respectively, where $\overline{a}_n := \sum_{i=1}^{p-1}\mathbf{1}_{\{\overline{d}_{n,i}>\overline{d}_{n,i+1}\}}$ and $\widehat{\alpha}_n := \sum_{i=1}^{p-1}\mathbf{1}_{\{\widehat{\delta}_{n,i}>\widehat{\delta}_{n,i+1}\}}$. Table 6 presents the results based on the simulations used to plot Figure 11.

p		30	50	100	150	200	300	400	500
n		90	150	300	450	600	900	1200	1500
Average % Violations	FSOPT	37.2	40.5	43.3	44.6	45.3	46.2	46.9	47.1
	QIS	34.6	33.6	31.0	28.9	27.5	25.9	24.4	23.6
Average Magnitude (%)	FSOPT	7.94	6.10	4.46	3.64	3.19	2.62	2.28	2.05
	QIS	3.15	1.89	0.92	0.59	0.41	0.26	0.19	0.14

Table 6: Summary statistics on order violations for FSOPT and QIS.

Interestingly, FSOPT has order violations in each one of the 4,950 Monte Carlo simulations generated to produce this table. For every pair of parameters (p, n), QIS has fewer order violations than FSOPT on average. The comparison is even more striking in terms of magnitude, as the order violations are systematically bigger for FSOPT than QIS, sometimes by an order of magnitude. Table 6 indicates that restoring order to QIS eigenvalues through the PAV algorithm is not a huge priority given that even the FSOPT, which cannot be beaten and assumes (unrealistic) foreknowledge of the population covariance matrix, tolerates more frequent and also larger order violations.

The overall trade-off is between, on the one hand, extra complexity (in moving from a closed-form solution to an iterative numerical algorithm) versus, on the other hand, a gain in accuracy so small as to be virtually negligible. Everybody is free to make their own decisions, but we believe that, on balance, the scales tip in favor of keeping the simple version of QIS without PAV.

F.2 Choice of the Kernel

In the main body of the paper, we select the Cauchy kernel for its simplicity, and also because its formula dovetails nicely with the classic work of Stein (1986). Equation (3.4) shows direct visual concordance between the *original* Stein shrinker and our own *smoothed* Stein shrinker. Statisticians who are familiar with the former will be comfortable with the latter. Not only is the formula built on the Cauchy kernel elegant, but its conjugate signal (required for handling Frobenius loss) is equally simple, cf. Proposition 4.4. The same cannot be said for other kernels.

Nonetheless, it is valuable as a sanity check to run a brief numerical comparison between various kernel choices. The default expectation is that they should have broadly similar performances. There is not much guidance from theory as to which kernel to choose in this context. The asymptotic theorems hold with a variety of standard kernels. An obvious family of alternatives are the kernels with compact support, which is well represented by the Epanechnikov (1969) kernel because it has some error-minimization properties, and has been used successfully by Ledoit and Wolf (2020). Another obvious alternative kernel is the Gaussian.

The first order of business is to adjust the shrinkage formulas to incorporate alternative kernels. This is mathematically tedious, so we will only report the shrinkers, leaving their conjugates and amplitudes to be worked out by the reader.

$$\text{Stein:} \qquad \widetilde{\theta}_n(x) = \frac{1}{p-1} \sum_{\substack{j=1 \\ \lambda_{n,j}^{-1} \neq x}}^{p} \lambda_{n,j}^{-1} \frac{1}{\lambda_{n,j}^{-1} - x} \tag{F.3}$$

$$\text{Cauchy:} \qquad \widehat{\theta}_n(x) = \frac{1}{p} \sum_{j=1}^{p} \lambda_{n,j}^{-1} \frac{\lambda_{n,j}^{-1} - x}{\left(\lambda_{n,j}^{-1} - x\right)^2 + h_n^2 \lambda_{n,j}^{-2}} \tag{F.4}$$

$$\begin{aligned} \text{Epanechnikov:} \qquad \widehat{\theta}_n^E(x) \coloneqq \frac{3}{p\, h_n} \sum_{j=1}^{p} \Bigg\{ & \frac{\lambda_{n,j}^{-1} - x}{10\, h_n \lambda_{n,j}^{-1}} + \frac{1}{4\sqrt{5}} \left[1 - \frac{\left(x - \lambda_{n,j}^{-1}\right)^2}{5 h_n^2 \lambda_{n,j}^{-2}} \right] \\ & \times \log \left| \frac{\sqrt{5} h_n \lambda_{n,j}^{-1} - x + \lambda_{n,j}^{-1}}{\sqrt{5} h_n \lambda_{n,j}^{-1} + x - \lambda_{n,j}^{-1}} \right| \Bigg\} \end{aligned} \tag{F.5}$$

Note that the expression between curly brackets is not defined when $|x - \lambda_{n,j}^{-1}| = \sqrt{5} h_n \lambda_{n,j}^{-1}$, so in this case it should be interpreted as just $\frac{\lambda_{n,j}^{-1} - x}{10\, h_n \lambda_{n,j}^{-1}}$.

$$\text{Gaussian:} \qquad \widehat{\theta}_n^G(x) = -\frac{\sqrt{2}}{p\, h_n} \sum_{j=1}^{p} \mathcal{D}\left(\frac{x - \lambda_{n,j}^{-1}}{\sqrt{2} h_n \lambda_{n,j}^{-1}} \right) , \tag{F.6}$$

where $\mathcal{D}$ denotes the *Dawson* integral: $\mathcal{D}(x) \coloneqq e^{-x^2} \int_0^x e^{t^2} dt$, a higher-transcendental function which is very slow to compute.

Once again, we re-run the convergence graph in Figure 5 but this time adding the Epanechnikov kernel. We also ran the Gaussian kernel, but the results were so close to Epanechnikov that it would have brought nothing of value to plot them, and only would have

reduced the legibility of the figure. So the results labeled "Epanechnikov" can actually be viewed as meaning "Epanechnikov and Gaussian" together. Figure 12 displays the results.

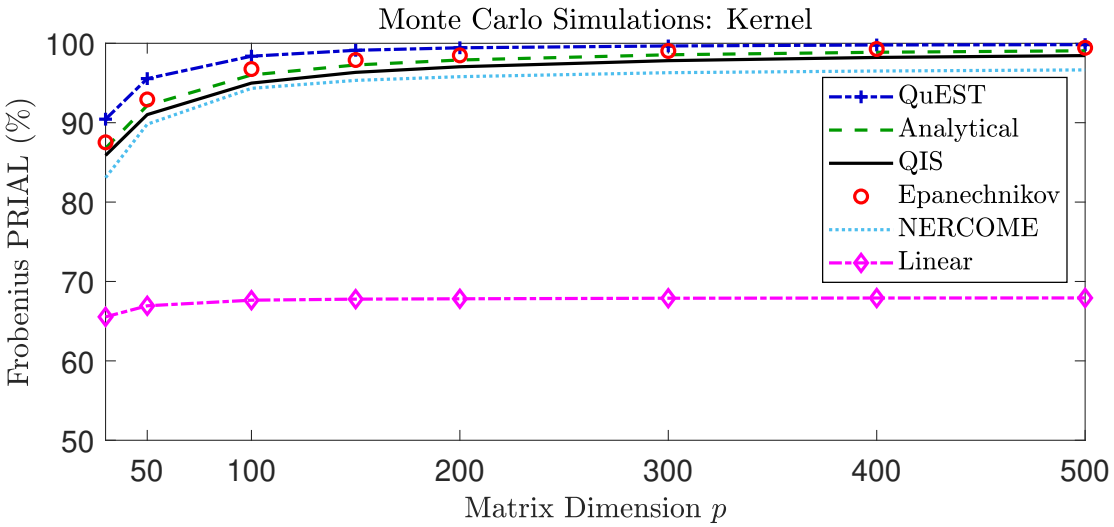

Figure 12: Evolution of the PRIAL as the matrix dimension and the sample size go to infinity together.

As expected, the results are very close. The performance gap between the estimators using different kernels is less than half the performance gap between the two numerical estimators (QuEST vs. NERCOME). This is very reassuring because it means that the method promoted here is robust against tweaks such as kernel modification. Although it is a little bit disappointing that the Cauchy kernel is not the best one in terms of performance (at least in this particular simulation), this is not why it was chosen in the first place. It was chosen for its close connection with Stein's nonlinear shrinkage formula and for its simplicity, as visual inspection of Equations (F.3)–(F.6) makes obvious.[7]

Another interesting piece of information to come out of Figure 12 is the comparison of the Analytical estimator (dashed line) with the QIS estimator smoothed by the Epanechnikov kernel (circles). They both use the same kernel, Epanechnikov, but differently. The former uses it to estimate the derivative of the c.d.f. of the spectral distribution of the sample covariance matrix, whereas the latter uses it to estimate the derivative of the first incomplete moment function of the spectral distribution of the sample precision matrix. Given that the results come out so close to each other, it indicates that the two approaches are essentially interchangeable for practical purposes. The main differences are in the mathematical proofs, and in the elegance of the resulting formulas. In this respect, we believe that the QIS approach (based on inverse eigenvalues) adheres more closely to the intrinsic nature of the mathematical problem at hand.

[7]It is unlikely that there exists a 'best' kernel across the board: indeed, Epanechnikov is slightly better than Gaussian in the lower dimensions ($p \leq 300$), whereas it is the reverse in the higher dimensions (400 and 500). Even if one embarked on a broader study of the question (which we shall not do here), the results would be purely experimental and contingent upon the specific setup of the Monte Carlo simulations, as there is no general theory to back them up.

F.3 Alternative Loss Functions

Even though the Frobenius loss is widely employed, the other two loss functions mentioned in Section 4 also have their respective advantages. In this section, we revisit the convergence graph of Figure 5, first by using the Inverse Stein's loss instead of Frobenius. Figure 13 displays the results.

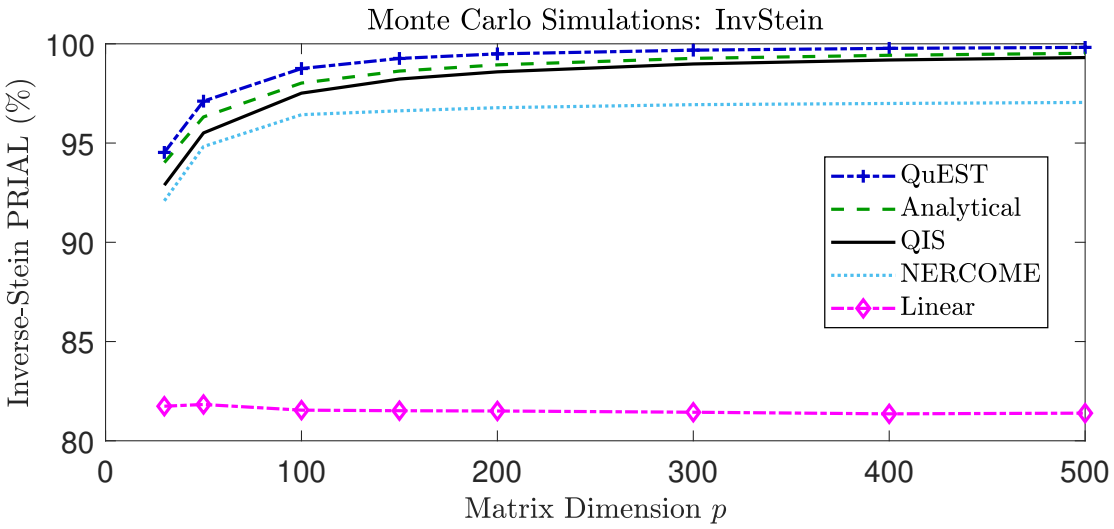

Figure 13: Evolution of the PRIAL as the matrix dimension and the sample size go to infinity together.

They are mostly analogous to Frobenius-loss results. If anything, Inverse Stein seems a bit more forgiving of the suboptimality of linear shrinkage (see the scale of the vertical axis). We then reconduct the same exercise, but this time with respect to the Minimum Variance loss of Definition 4.3. Figure 14 displays the results.

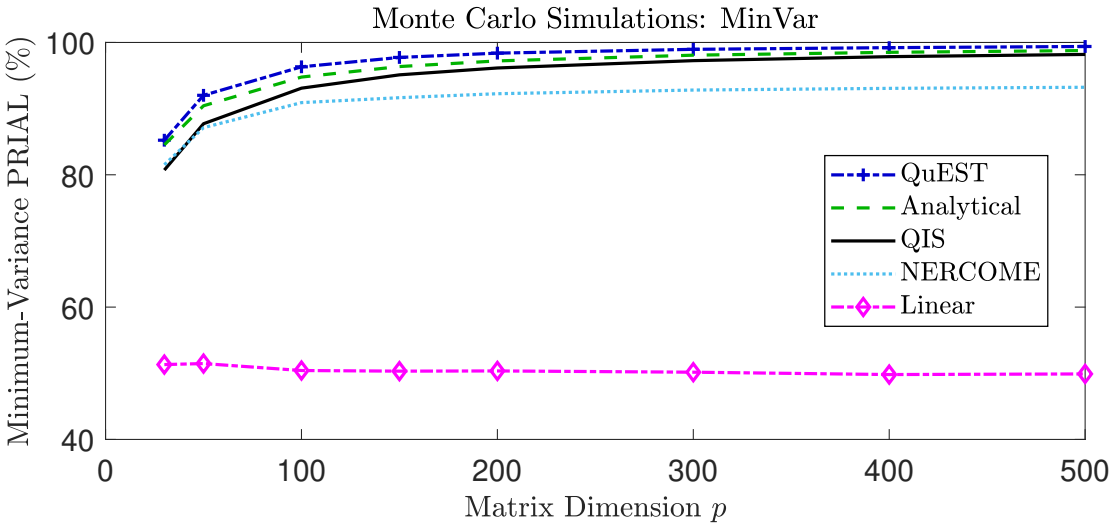

Figure 14: Evolution of the PRIAL as the matrix dimension and the sample size go to infinity together.

One can see that there is essentially no difference in terms of numerical performance, relative the corresponding simulations under Frobenius loss. This is what we expected based on the theoretical results of Corollary 4.1.

F.4 Smoothing Hyperparameter

In Section 6.2, we presented the surface of optimal h_n's as a function of (p, n). Each point in Figure 4 was the average of 12,000 numbers: 6 population spectra $\times$ 2 loss functions $\times$ 1,000 simulations. It is worthwhile checking whether the inverted V-shaped pattern remains the same when we partial out the simulations by shape of population spectrum. In this exercise, we will obtain 6 different optimal smoothing parameter surfaces, and each point in each figure will be the average of 'only' 2000 simulations. Figure 15 displays the results.

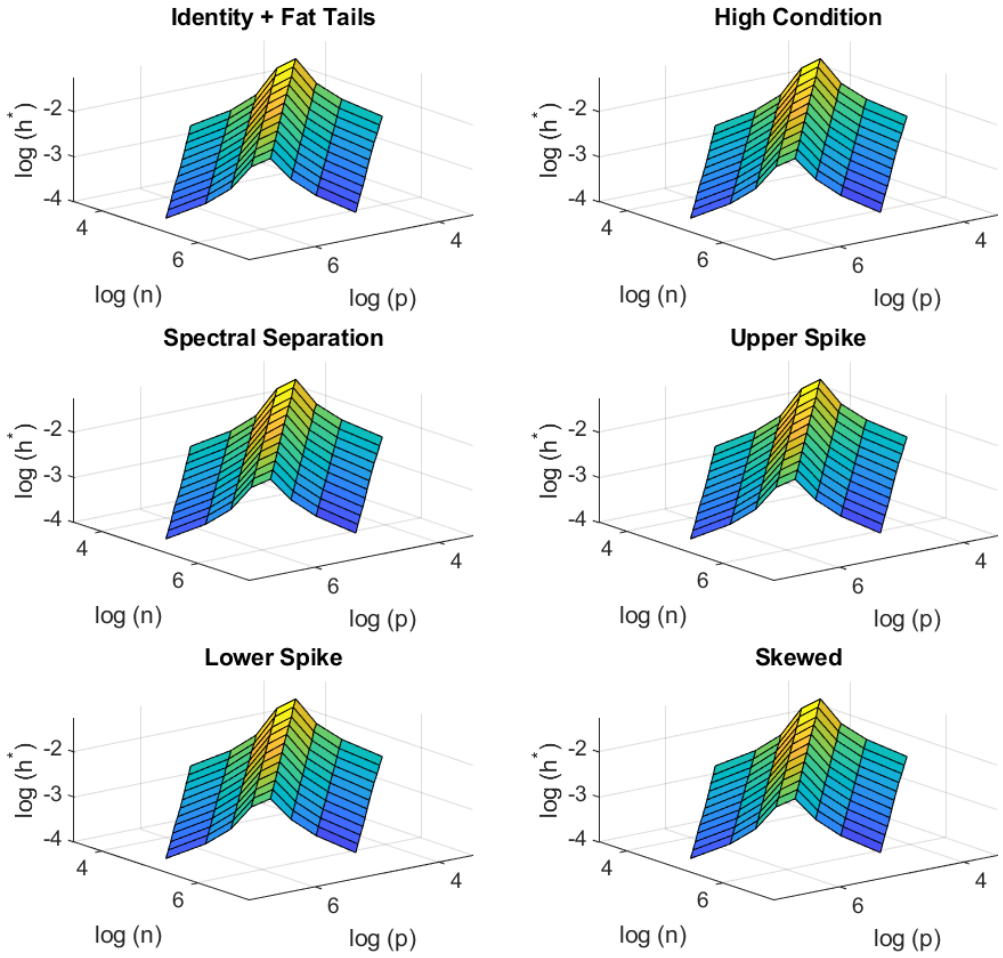

Figure 15: Optimal smoothing parameter as a function of the matrix dimension and the sample size for various shapes of the population spectrum.

The consistency across scenarios demonstrates the robustness of our choice of smoothing

parameter. In order to further explore this question, we fit the linear regression model of Equation (6.2) separately for each population-spectrum shape. Table 7 presents the results.

Population Spectrum	Identity+ Fat Tails	High Condition	Spectral Separation	Upper Spike	Lower Spike	Skewed
Intercept	0.118	0.444	−0.214	0.316	0.252	0.620
	(0.082)	(0.073)	(0.108)	(0.078)	(0.079)	(0.074)
$\log\big[\min(p/n, n/p)\big]$	0.757	0.623	0.662	0.669	0.671	0.574
	(0.020)	(0.018)	(0.026)	(0.019)	(0.020)	(0.018)
$\log[p]$	−0.378	−0.421	−0.306	−0.404	−0.394	−0.447
	(0.015)	(0.013)	(0.020)	(0.014)	(0.015)	(0.013)
n	192	192	192	192	192	192
R^2	0.914	0.919	0.816	0.913	0.909	0.916

Table 7: Linear regression fits of model (6.2) to the optimal bandwidth in log-space.

The results are very similar across population spectra. The coefficient on $\log\big[\min(p/n, n/p)\big]$ is close to 0.70, give or take some, and the coefficient on $\log[p]$ is close to -0.35.

Also of interest is the extent to which the performance of the QIS estimator is affected if we change the formula for the smoothing parameter. We consider a general specification:

$$h_n := \kappa \times \min\left(\frac{p^2}{n^2}, \frac{n^2}{p^2}\right)^{\alpha} \times \ p^{-\beta} \ . \tag{F.7}$$

The formula promoted in Section 6.3 corresponds to the particular case $\kappa = 1$, $\alpha = \beta = 0.35$. In order to measure performance, we start from the baseline scenario and vary p and n as per Section 6.2, then average the Frobenius PRIALs. Since there are $8 \times 12 = 96$ pairs of (p, n) parameters, and we run 1,000 Monte Carlo simulations for each (p, n), each cell in Table 8 below represents and average of $96,000$ simulations.

$\beta = 0.30$	$\alpha = 0.30$	$\alpha = 0.35$	$\alpha = 0.40$
$\kappa = 0.5$	96.6	96.5	96.3
$\kappa = 1$	96.1	96.3	96.5
$\kappa = 2$	92.9	93.6	94.1

$\beta = 0.35$	$\alpha = 0.30$	$\alpha = 0.35$	$\alpha = 0.40$
$\kappa = 0.5$	96.2	96.0	95.6
$\kappa = 1$	96.5	96.7	96.7
$\kappa = 2$	94.3	94.8	95.3

$\beta = 0.40$	$\alpha = 0.30$	$\alpha = 0.35$	$\alpha = 0.40$
$\kappa = 0.5$	95.5	95.1	94.5
$\kappa = 1$	96.7	96.7	96.7
$\kappa = 2$	95.4	95.8	96.0

Table 8: Average Frobenius PRIAL (in %) for 27 variants of the smoothing parameter.

Five different specifications, including our default one, achieve the best score of 96.7%. The only universal pattern that emerges is that $\kappa = 2$ is always worse than $\kappa = 1$. Apart from that, the results are distributed irregularly. 22 out of 27 configurations fall within 1% of the overall mean (which is 95.7%). The worst performance is 92.9%, still a honorable PRIAL. On this basis, we can safely conclude that performance is not sensitive to the choice of the smoothing hyperparameter, as long as it remains generally in the same neighborhood as the formula adopted in Equation (6.3).

F.5 Comparison With Stein's Original Estimator

Given that the Stein (1986) estimator plays an important part as starting point of our investigation, it would be interesting to see how it performs compared to the two estimators that we have obtained: the Linear-Inverse Shrinkage (LIS) estimator of Equation (3.5), and the Quadratic-Inverse Shrinkage (QIS) estimator of Equation (4.16). Indeed, one can consider these three estimators a family that stands apart from other strands of the literature on covariance matrix estimation: Going from Stein's original to LIS is just a matter of inverting the order of the shrinkage step and the regularization step; which then enables us to go to QIS in order to deal with loss functions other than Stein's loss, bring the amplitude into play, and address the $p > n$ case.

Stein, LIS, and QIS are optimized with respect to different loss functions, so it is only fair to compare them with respect to all the loss functions mentioned in the main body of the paper: Stein's loss, Inverse Stein's loss, Frobenius loss, and Minimum Variance loss. So there are 3 estimators and 4 loss functions, yielding $3 \times 4 = 12$ different PRIALs. Given that this number is starting to get unwieldy, and that this exercise is somewhat derivative to the main thrust of the paper, we content ourselves with looking at the baseline scenario of Section 7.1, which should be representative of the broader patterns. Table 9 presents the results, each cell representing an average of 500 simulations.

Prial (%)	Stein's loss	Inverse Stein's loss	Frobenius loss	Minimum Variance loss
Stein	97.9	84.5	78.5	95.2
LIS	96.4	87.4	82.6	93.4
QIS	66.5	98.6	97.1	96.2

Table 9: Comparing three estimators with respect to four different loss functions.

As expected, Stein's original estimator and the LIS estimator do well with respect to Stein's loss, since this is the loss function with respect to which they are optimized; and they do less

well with respect to the other three loss functions. Conversely, the QIS estimator does well with respect to Inverse Stein's loss, Frobenius loss, and Minimum Variance loss, since it is optimized with respect to all three; and less well with respect to Stein's loss.

As a secondary observation, there is no strict ordering between Stein's estimator and the LIS estimator: it depends on the particular loss function. One can also see that, whereas the three loss functions that belong in the same group (Inverse Stein's loss, Frobenius loss, and Minimum Variance loss) treat the QIS estimator in very much the same way, there is considerable variation between them on how lenient they are with respect to estimators optimized for Stein's loss. For example, Stein's original estimator scores a markedly lower (but still respectable) PRIAL of 78.5% with respect to Frobenius loss, and a much higher one of 95.2% with respect to the Minimum Variance loss. Overall, the Minimum Variance loss seems the most tolerant of nonlinear shrinkage estimators that are optimized with respect to the 'wrong' loss function. By contrast, in Section F.3, it was the Inverse Stein's loss that was most forgiving of linear shrinkage.

From this investigation, we can conclude that the original estimator of Stein (1986) holds its own surprisingly well. By working to improve upon it, we are taking on not a minnow but a giant. In particular, it is hard to promote the LIS estimator over Stein's original one, since the inversion of order between the nonlinear shrinkage step and the numerical regularization step (isotonization for Stein, kernel smoothing for LIS) does not make enough of a difference in practice. The main advantage is theoretical in the sense that LIS can prove optimality by harnessing the power of large-dimensional asymptotics. This is to a large extent why this paper has not promoted the LIS estimator as strongly as the QIS estimator. Speaking of which, Table 9 shows that it makes much difference if one employs an estimator optimized with respect to the 'right' loss function (as determined by the problem at hand). It would be hard to recommend the Stein (1986) estimator for anyone interested in the Inverse Stein's loss, the Frobenius loss, or the Minimum Variance loss. In all such cases, the QIS estimator is more suitable, as we do not know *a priori* to what extent the Stein estimator will underperform.